A Thesis, Allegedly

by

Jisoo Hong

S.B., Massachusetts Institute of Technology (2018)

Submitted to the Institute of Data, Systems, and Society
in partial fulfillment of the requirements for the degree of

Master of Science in Technology and Policy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 2022

Author . . . .
Institute of Data, Systems, and Society
May 6, 2022

Certified by . . . .
Graham M. Jones
Professor of Anthropology
Thesis Supervisor

Accepted by . . . .
Noelle Eckley Selin
Professor, Institute for Data, Systems, and Society and
Department of Earth, Atmospheric and Planetary Sciences
Director, Technology and Policy Program

A Thesis, Allegedly

by

Jisoo Hong

Submitted to the Institute of Data, Systems, and Society
on May 6, 2022, in partial fulfillment of the
requirements for the degree of
Master of Science in Technology and Policy

Abstract

Current approaches for studying the spread of misinformation on social media tend to focus on the factual integrity of shared content and the reach or circulation of false claims. However, a focus on the truth value of content can obscure the embeddedness of information in social, communicative practices. One way of apprehending the sociocultural dimensions is through an analysis of the stances people take toward the information they circulate online.

In this thesis, we investigate how language mediates perceptions of truth and reality through a close examination of how data is animated as evidence. This process, we argue, is fundamentally interactional and dialogic. Using sociolinguistic and natural language processing (NLP) techniques, we demonstrate how specific features of evidential talk, such as the use of epistemic adverbs like *allegedly* or *supposedly*, can dramatically alter how evidence is taken up in discussions of scientific controversy. We present the *hearsay effect*, a numerical measure mapping the entextualization of data as hearsay to its engagement and circulation on social media, to characterize how subtle inflections in epistemic modulation shape the social life of evidence. We find that the hearsay effect is variably salient in different discursive communities, and is particularly prominent in our case study of evidential discourse amongst ufologists on Twitter. We suggest that this analysis of the strength of weak evidence within contested sites of knowledge production offers new ways of conceptualizing how information and misinformation is animated in the online public sphere.

Thesis Supervisor: Graham M. Jones
Title: Professor of Anthropology

Acknowledgments

I have been extremely lucky throughout my MIT career to have crossed paths with many wonderful people, whose care and support have made this thesis possible.

My advisor, Graham, whose kindness, encouragement, and creativity shaped not only this work but my approach to research and beyond. Whenever it felt like I was grasping at straws, our conversations and whiteboarding sessions would never fail to generate new perspectives and ideas and reinspire me. I am endlessly thankful for his generosity and trust.

My collaborators, Maya, Bambi, and Emma, who lent their time and experience to shed invaluable light on this project, as well as Arvind and the Language and Technology Lab for their advice and feedback.

The Richard de Neufville Fellowship and MIT SHASS Fang Fund, for believing in and funding this work.

Barb, Elena, Frank, and Noelle, who were the first to welcome us to TPP and who supported us throughout our academic journeys, both on Zoom and in person.

My friends, for keeping me afloat with their long calls, visits, cooking, and spontaneous adventures, and for making Cambridge such a warm place to call home.

My parents, who have worked tirelessly to make my world bigger and brighter from the very beginning, for their continual love and support.

This project would not have taken flight if not for these people. Thank you.

Contents

List of Figures

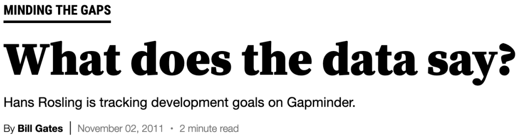
List of Tables

Chapter 1

Can the data speak for itself?

Throughout the coronavirus pandemic, data has occupied a conspicuous and hotly contested place in the public imagination, casting new light on how constructions of truth and reality are negotiated on social media. Dashboards tracking data on case counts and hospitalization rates have become crucial to the project of informing the public and rationalizing measures such as school closures and mask mandates. At the same time, anti-maskers and Covid-denialists have also been hard at work producing and circulating "counter-visualizations" – often using the same datasets as health officials – to advocate for radically different policy changes (Lee et al. 2021). By urging community members to "follow the data," these groups mobilize orthodox data practices to generate evidence for unorthodox scientific claims. But can the data really speak for itself?

The ways in which this expression is implemented are manifold and worth considering in their own right. "What does the data say?" Bill Gates asked in 2011, promoting initiatives where data take on a more capacious role in tracking global development and the impact of aid programs (1-1). This 4/20, the data activist collective data4blacklives shared "What the Data Says" on its Instagram, looking at the impact of the legalization of cannabis on Black people across the US (1-2). The phrase also reverberated across Twitter throughout the pandemic as people struggled to make sense of conflicting evidence online; Florida governor Ron DeSantis famously boosted a controversial article on the efficacy of mask mandates in a tweet while



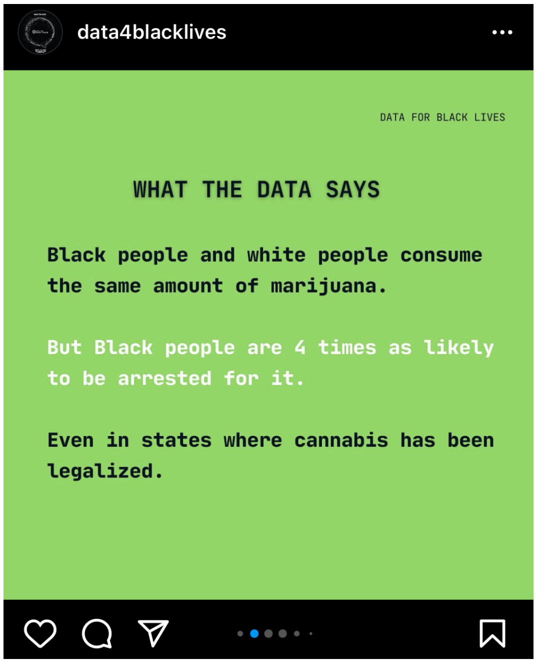
MINDING THE GAPS

What does the data say?

Hans Rosling is tracking development goals on Gapminder.

By Bill Gates | November 02, 2011 • 2 minute read

Figure 1-1: Post from Bill Gates' blog, GatesNotes



data4blacklives

DATA FOR BLACK LIVES

WHAT THE DATA SAYS

Black people and white people consume the same amount of marijuana.

But Black people are 4 times as likely to be arrested for it.

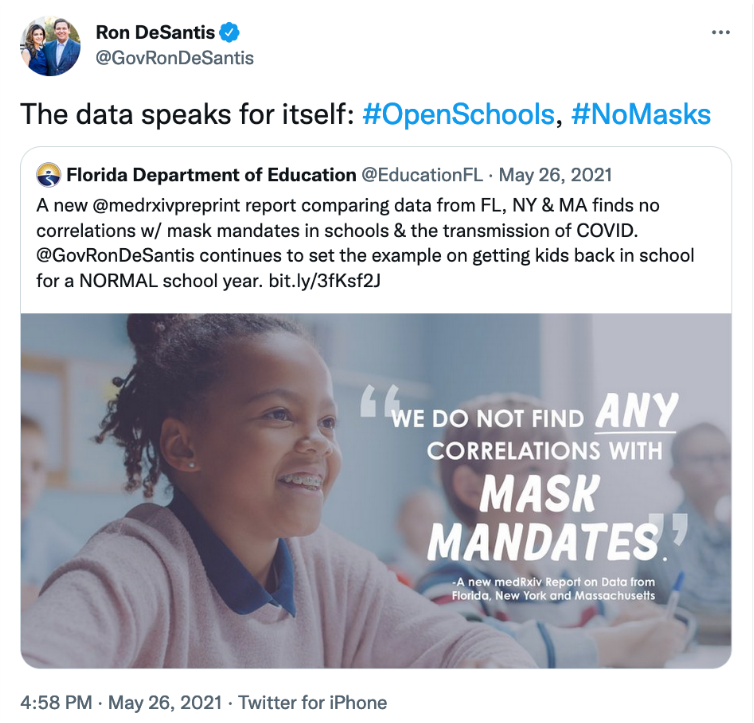
Even in states where cannabis has been legalized.

Figure 1-2: Instagram post by @data4blacklives

proclaiming that "the data speaks for itself" (1-3) last May.

Each of these examples personifies "data" as an independent speaking subject, but it is clear that someone *else* (the entity sending the tweet) is speaking with or for the data. Indeed, the data cannot ever speak for itself – because it must be animated and advocated for, and then accepted by others, before it can claim meaning. This thesis contends that a focus on the way that data is *entextualized* – that is, turned into a communicative message (Bauman and Briggs 1990) – helps to surface these dynamics, which are so often glossed over with rhetorical sleight of hand. By focusing on how data is invoked in conversation to shape public perceptions of truth and reality, we shed new light on how data comes to develop a social life of its own.

The DeSantis tweet is an illustrative example of the social relations that animate data, partly because the interactional affordances of Twitter bring them to the fore.



Ron DeSantis
@GovRonDeSantis

The data speaks for itself: #OpenSchools, #NoMasks

Florida Department of Education @EducationFL · May 26, 2021
A new @medrxivpreprint report comparing data from FL, NY & MA finds no correlations w/ mask mandates in schools & the transmission of COVID. @GovRonDeSantis continues to set the example on getting kids back in school for a NORMAL school year. bit.ly/3fKsf2J

"WE DO NOT FIND ANY
CORRELATIONS WITH
MASK
MANDATES."
-A new medRxiv Report on Data from Florida, New York and Massachusetts

4:58 PM · May 26, 2021 · Twitter for iPhone

Figure 1-3: Tweet by Ron DeSantis

In the tweet, DeSantis quotes the Florida Department of Education, which links an article published on medrxiv (a preprint server for medical research) reporting that mask mandates are not correlated with the transmission of Covid in schools, while simultaneously praising the governor's stance on school reopenings. The study compares outcomes from Florida and two other states using a dataset collected by economist Emily Oster and collaborators at the Covid-19 School Response Dashboard project in order to make its claims. At this point, the data has already been "spoken for" by multiple actors: the people (probably state bureaucrats) who published it online; the academics who aggregated it, ran the regressions to interpret it, and wrote a paper about it; the Florida Department of Education, who animated it on Twitter by aligning it with DeSantis' policy agenda; and DeSantis, who *re*-animated it on Twitter, as did the other 1,618 accounts that retweeted it and 147 accounts that quoted it.

Despite this ostensible show of support, this data is received with a great deal of scrutiny. Many of the tweets in the replies condemn DeSantis for implying that the Florida data is credible to begin with: "Flawed numbers. We know how you roll." (larryca66028461); "His data isn't the true data" (Madmom42004988); and "Oh I know dear especially since you covered up the real data." (IndiaJenkins1). Others attack the quality of the paper: "not peer reviewed buddy. did you go to college or pay someone to do it for you. you gotta love men willing to risk children's health." (live4literacy). Still others criticize his interpretation of the data: "I love the picking and choosing of data when it suits people" (KitCMiller2); and "The data says THE EXACT OPPOSITE OF THAT, you absolute human garbage sputtering moron. Why do you insist on forcing unvaccinated children to be unprotected when your state has consistently had the HIGHEST PEDIATRIC POSITIVITY RATE in the country since the pandemic began?" (EcoSexuality).

The critical register in these replies makes plain that the social life of data – whether it is taken up by the public as evidence, or not – is not just determined by the veracity of the information at hand. @KitCMiller2, @EcoSexuality, and others are clearly very aware that politicians regularly use and manipulate numbers for their own political gain. In articulating their own stances, these users introduce new interpretations of the data and attempt to seize its evidentiary status. This in turn indelibly shapes the landscape of participation and possibility for future interlocutors. Truth, then, is not so much something data can "capture" for once and for all, but rather something to be negotiated through it. With this in mind, we seek to apprehend the sociocultural dimensions through an analysis of the linguistic dimensions of evidentiality.

The thesis is divided conceptually into three main parts. It begins with a theoretical exploration of how data becomes evidence. To this end, Chapter 2 draws upon literature on the subjectivity of data and evidence from science and technology studies (STS) and sociolinguistics, arriving at *stance* as a useful mechanism with which to bridge the two. We review how stance is theorized and operationalized in linguistic anthropology and computer science, and describe challenges in translating an anthropologically-meaningful concept of stance into something that is computable. To this end, we discuss categorizations of linguistic features that *index* stance, called stance markers, in order to gain analytical traction on the problem. Chapter 3 studies the deployment of these stance markers within a discourse community in which truth, fact, and the legitimacy of data are continuously disputed: #ufotwitter. With the Twitter v2 API, we develop a novel dataset of tweets using the hashtag to discuss emerging evidence in the field, and detail results from a digital ethnography (or hashtag ethnography, cf. Bonilla and Rosa 2015) of the #ufotwitter community. The chapter also records our efforts in measuring variation in how stance is mediated across the corpus with respect to differences in social status, interactional patterns, substantive topics, participatory inflection, and rates of engagement. We find that the use of the stance marker *allegedly* in the #ufotwitter corpus is related to statistically outsized levels of engagement and circulation. Chapter 4 characterizes this finding as the *hearsay effect*. The chapter provides some theoretical background as to why *allegedly* is specifically indexical of *hearsay*, and evaluates the robustness of the hearsay effect in terms of causality and generalizability using a mixed-methods approach. We find that the hearsay effect becomes particularly salient in discursive spaces where participants: (1) are collectively preoccupied with the problem of evidential production; and (2) epistemic or institutional authority is absent. We conclude with a discussion of what this might mean for information ecologies on social media, where data is circulated as evidence to variable perlocutionary effect.

Chapter 2

What we talk about when we talk about evidence

If the data cannot speak for itself, how else can we account for its rhetorical function? In this chapter, we distinguish between data as *information* and data as *evidence*, contending that the latter only comes into being when the data is put forth to support an argument or a point of view. With *stance* as a point of departure, we search for a method to assess how language mediates its uptake as constructions of truth and reality, offering new perspectives for apprehending the interactional and communicative embeddedness of data.

2.1 All data is perspectival

Data is often rendered in the technocratic imagination as something that is cheap, abundant, and waiting to be unearthed from the "ground" of social life[1] (Puschmann and Burgess 2014; Hwang and Levy 2015; Mejias and Couldry 2019). Despite the pervasiveness of this narrative, critical scholars have worked to draw attention to the pernicious effects of obfuscating the subjectivity of data, and have developed theoretical accounts of data as fundamentally social relations (D'Ignazio and Klein

[1] Though companies like Google have pushed back on criticism of slogans such as "data is the new oil" with assertions that data "is more like sunlight than oil" – a replenishable, ownerless resource that can be harvested sustainably for the collective benefit of society (Couldry and Mejias 2019)

2020; Viljoen 2021). Indeed, all data is perspectival, an idea that proceeds directly from the work of anthropologist Charles Goodwin, who wrote that "all vision is perspectival and lodged within endogenous communities of practice" (Goodwin 1994). There is nothing raw about data, as Travis D. Williams reminds us in *Raw Data is an Oxymoron* (Gitelman 2013). Whether it is done consciously or subconsciously, data is always produced, processed, and consumed with a social objective or purpose in mind. As such, data is always dependent on a social actor to collect it, make sense of it, contextualize it, and translate it (Leite and Mutlu 2017). It takes labor to transform data into capital (Sadowski 2019), and into claims of fact and truth.

This thesis proposes that the work that undergirds this second transformation is *evidential* in nature. To better make sense of this proposition, it is useful to be fastidious with some terminology that data is frequently conflated with: namely *information* and *evidence*. Following sociologist Howard Becker (2017), we interpret data to be a preserved record of information until it is mobilized in support of an idea, or argument. It is through this process of *becoming evidence* that data is scrutinized as a potential representation of reality. The evidence may then be accepted as fact or truth, with *accepted* being the operative word: the evidence must convince an audience of its validity, and of its weight (Becker 2017).

What gives data-as-evidence its persuasive power? Much of the extant scholarship approaches this question in terms of the authority that is granted to the data exogenously. These works consider why and when people find data to be credible ways of knowing and communicating. In outlining a sociology of quantification, Espeland and Stevens (2008) organize the sources of this authority into four broad categories: its perceived accuracy or validity as a proxy for reality (Anderson and Fienberg 1999; Desrosières 2001); its usefulness in solving problems (Carson 2007; Didier 2002; Porter 1995); its ability to accumulate and link users who are invested in the data itself (Feldman and March 1981; Kalthoff 2005; Latour 1988; Callon 1984); and its long and evolving relationship with ideals of rationality and objectivity (Daston 1992; Nussbaum 1986; Weber 1978). However, not much has been written about how people use language to mediate the credibility and authority of data in conversation – in their "endogenous communities of practice" (cf. Goodwin 1994). In this thesis, we seek to understand how language shapes the process of making data into evidence through an analysis of the stances people take toward the information they circulate online.

2.1.1 Stance, encoded

Stance does not have one set theoretical definition in sociolinguistics and linguistic anthropology, but broadly refers to the marking of attitudinal or ideological perspective – or how people position themselves in conversation (Biber 2006; Beach and Anson 1992; Kiesling 2022). Indeed, this concept has been studied under many different labels (Biber 2006; Chindamo, Allwood, and Ahlsen 2012). Jones and Harris (1967) examine positions of speech (pro, anti, or equivocal) in terms of *attitude*; Martin (2000) discusses intersubjective and ideological positioning through the lens of *appraisal*; Hunston and Thompson (2000) consider *attitude* and *stance* under the broader umbrella of *evaluation*. Even within stance, there are subdivisions. Biber et al.(1999) differentiate between *epistemic* and *affective* stance. In this framework, *epistemic* stance deals with certainty, doubt, actuality, source of knowledge, imprecision, viewpoint, and limitation, while *affect* is more concerned with states, evaluations, emotions, and attitudes. Other categorizations include *social*, *authorial*, *interpersonal*, and *intentional* stance (Kockelman 2004). Social stance might refer to the "ethical horizons of a group relative to language, thereby implicating stance in linguistic ideologies." *Authorial* stance might refer to the ideology of authors with respect to how they position themselves in and to a text. *Interpersonal* stance might be studied to analyze the particular perspectives of participants in a speech event, and *intentional* stance might be of interest to apprehend social behavior in terms of putative mental states such as belief, desire, and fear.

Efforts to catalogue how stance is encoded are wide-ranging. Within the scope of studies on the English language, different methodologies have been employed to identify the lexical, modal, and grammatical means by which stance is indexed. Some studies specifically focus the discussion on the function of adverbials in marking stance. For example, Biber and Finegan (1988) analyze stance via the distribution of a particular class of stance-encoding words and phrases, *stance adverbials* – adverbs that indicate attitudes, feelings, judgments, or commitment concerning the message. Hyland (1996) studies stance-taking in the academic research context by studying the range of functions and grammatical devices used to express tentativeness and possibility in research articles. This work shows how the mobilization of adverbs such as *generally*, *approximately*, *partially*, or *possibly* can index the speaker's attitude towards the accuracy of a proposition. Quirk et al. (1985) also discusses adverbial stance markers in terms of *disjuncts*, distinguishing between *style* disjuncts – adverbs such as *truthfully*, *seriously*, *bluntly*, *approximately*, or *briefly*, indicating the manner of speech – and *content* disjuncts – adverbs such as *certainly*, *doubtfully*, *amazingly*, *annoyingly*, or *fortunately*, indicating the speaker's certainty about or judgment of the content.

Other works have turned their attention to different markers of stance in English – for example, modals such as *might* or *should*, or "opinion" and "perception" verbs such as *believe*, *think*, *seem*, and *disappear* (Biber and Finegan 1988). Kärkkäinen (2003) focuses on the salience of the syntactic/pragmatic phrase *I think* as a measure of epistemic stance-taking in conversational English. Studies from Thompson and Ye (1991) and Hyland (2002) also describe how "reporting" verbs like *state*, *consider*, and *find* can function as expressions of different kinds of stance meanings.

Throughout this kaleidoscopic array of scholarship runs a common thread, highlighting the reason stance is so meaningful in linguistics. Taken as a whole, scholars apprehend stance as the performance of *positionality* – how speakers and writers are necessarily engaged in positioning themselves vis-à-vis their words and texts (which are embedded in their own histories of linguistic and textual production), their interlocutors and audiences (whether they are actual, virtual, projected, or imagined), and with respect to a context that they simultaneously respond to and construct linguistically (Jaffe 2009). In other words, stance is inherently subjective, dialogic, and interactional (Goodwin and Goodwin 2005; Kockelman 2004; Du Bois 2007).

This understanding of stance is analytically rich. However, in order to understand how stance manifests on social media, where talk is generated prolifically and parallelly, we need a method to scale stance to hundreds of thousands or millions of utterances. Because stance-taking is such a multi-dimensional concept, we begin to break down how stance is operationalized on different interactional axes with a review of two frameworks from the linguistic anthropological literature.

2.1.2 Stance, operationalized

In this section we review two frameworks from linguistic anthropology describing how the social elements of stance can be parsed and disambiguated.

Kockelman (2004) draws from sociologist Erving Goffman's definition of the role of the speaker (1981) and linguist Roman Jakobson's conceptualization of grammatical categories in terms of events (Jakobson, Waugh, and Monville-Burston 1990) in order to theorize the semiotic and linguistic realization of stance. Specifically, he observes that Goffman's decomposition of the role of the speaker into that of the *animator* (the one speaking the words), *author* (the one composing the words said), and *principal* (the one who is committed to what the words say) can be combined with Jakobson's description of the *speech event* (the world in which speaking occurs), *narrated event* (the world spoken about), and *narrated speech event* (a spoken-about world in which speaking occurs) to characterize stance-taking as a set of triadic relations. Stance, he argues, is a commitment to a construal of an event, whereby someone (the principal) commits to a *narrated event* (the realm of figures) relative to a *speech event* (the realm of animators). This disambiguation of roles and events in stance-taking is realized through a number of linguistic resources – from grammatical categories such as mood and status (or epistemic modality) to lexical categories such as complement-taking predicates (including *believe*, *want*, and *fear*).

Meanwhile, Du Bois (2007) defines stance as "a public act by a social actor," and offers the *stance triangle* as another useful model for thinking about stance. Under this model, the *stance act*[2] is understood to be three acts in one: two sets of evaluation

[2] The stance act is usually the evaluation of a *stance object* (which could be a "real" object or simply a figure in the discourse).

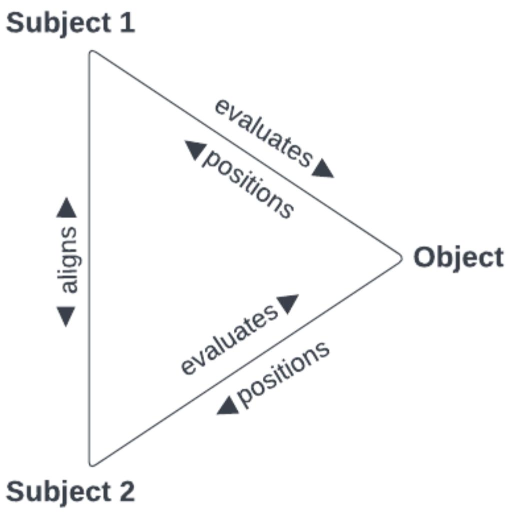


Figure 2-1: Stance triangle (Du Bois 2007)

of the object (alternatively, two instances of taking a position on the object), and the relative alignment or misalignment created by the two evaluation/positioning moves. A graphic representation of these relationships is shown in 2-1.

Du Bois gives the following as a simple example of a stance act:

(1) SAM: I don't like those.
ANGELA: I don't either

Here, Sam gives a negative evaluation of whatever is referenced in *those* (let's say they are shopping for chairs, so Sam is a subject and a chair is the object). Angela (another subject) also gives a negative evaluation of the chair, thereby aligning herself with Sam. Of course, not all discourse is so neat. Sometimes, the stance object is implied and difficult for an external observer to infer. Even when the stance-taker and the stance object are easily identifiable, the stance that the stance-taker is responding to may require the recursive analysis of many stance acts, or may not be discernible at all. (Consider, for example, a variation of the above in which Sam and Angela are being sarcastic.)

These operationalizations of stance are complex and multi-dimensional, and more suited for qualitative – rather than quantitative – analysis. Even though one of the stance triangle's innovations is to conceptualize the alignment between participants in a dialogue as a scalar value (rather than a binary state of alignment or misalignment), it is not immediately obvious how to assign the scalar value from the other states of evaluation and positioning, making it more useful symbolically, as a signal of nuance or gradation, or for small-scale analysis. Kockelman's characterization of stance as a commitment to a construed event is similarly difficult to translate into a scalable instrument or metric. To apprehend the articulation of stance in large social media corpora such as the ones studied in this thesis, computational methods must also be considered.

2.1.3 Stance, computed

Stance detection in the computational literature is generally formulated as a classification problem. As input, a piece of text and a *target* (an entity, concept, event, idea, opinion, claim, topic, etc) are given. The author's stance is predicted as the output, usually from a category label from this set: {*Favor*, *Against*, (*Neither*)} (Küçük and Can 2021). Work to automatically identify stance from text is also sometimes called stance identification (Zhang et al. 2017), stance prediction (Qiu et al. 2015, debate-side classification (Anand et al. 2011), debate stance classification (Hasan and Ng 2013), rumour stance classification (Zubiaga et al. 2018), or fake news stance detection (Pomerleau and Rao 2017). There is considerable heterogeneity in the categorization of stance as an output label across these studies. For example, rumour stance classification uses the labels {*Supporting*, *Denying*, *Querying*, *Commenting*}, Fake news stance detection uses the labels {*Agrees*, *Disagrees*, *Discusses* (the same topic), *Unrelated*}. Simaki et al. (2018) use six stance categories {*Contrariety*, *Hypotheticality*, *Necessity*, *Prediction*, *Source of Knowledge*, *Uncertainty*} in their work.

Stance detection has many potentially-profitable applications ranging from social media (fact-checking, fake news detection, content moderation, etc.) to public opinion mining (for political campaigns, advertising, and marketing). As such, a wide array of computational techniques have been employed in an effort to solve this problem. The methods generally fall within one of three buckets: feature-based machine learning, deep learning, and ensemble learning (Küçük and Can 2021). Support vector machines (SVMs) (Hacohen-Kerner, Ido, and Ya'akobov 2017; Mohammad, Sobhani, and Kiritchenko 2017; Tsakalidis et al. 2018) and logistic regression (Ferreira and Vlachos 2016; Zhang et al. 2017), for example, are sometimes used with handcrafted features (character and word ngrams, part-of-speech tags, hashtags, sentiment dictionaries, etc) for classification. Deep learning approaches use long short-term memory (LSTM) architectures (Augenstein et al. 2016; Dey, Shrivastava, and Kaushik 2018; Wei, Lin, and Mao 2018) and convolutional neural networks (CNNs) (Wei et al. 2016; Zhou, Cristea, and Shi 2017) for both feature extraction and end-to-end classification. Some models use attention mechanisms to improve performance (Dey, Shrivastava, and Kaushik 2018; Sobhani, Inkpen, and Zhu 2017; Wei, Lin, and Mao 2018; Zhou, Cristea, and Shi 2017). Ensemble learning methods range from simple random forest schemes (Tsakalidis et al. 2018) to more complex algorithms implementing semi-supervised user modeling (Fraisier et al. 2018) or combinations of deep learning architectures (Zhang et al. 2017).

Many of these systems are highly experimental and do not yet give robust results. Even with expensive deep learning techniques, the top-performing models submitted for SemEval-2016[3] do not outperform the baseline system provided by the competition organizers, which uses SVM with handcrafted features (with an F-score of 68.98).

However, the formulation of stance detection as a problem of predicting a binary (or ternary) label {*Favor*, *Against*, (*Neither*)} arguably dooms it from the start from effectively capturing the inherently dialogic and interactional elements of stance-taking.

Two studies, Pavalanthan et al. (2017) and Kiesling et al. (2018), make this insight and attempt to operationalize a sociolinguistic construct of stance by introducing new correspondences between *stance dimensions* and the lexical features that characterize them. Given that the literature on computational approaches for detecting interactional stance is extremely limited, this thesis takes particular interest in how these stance correspondences are defined.

First, Pavalanthan et al. (2017) aim to identify latent stance dimensions in a

[3] A competition for stance detection in English tweets (Mohammad et al. 2016)

custom Reddit corpus. To begin, they compile a lexicon of stance markers indexical of interactional stance-taking, starting from a seed lexicon of stance markers from Biber and Finegan (1989). This list includes certainty adverbials (*actually*, *of course*, *in fact*), affect markers (*amazing*, *thankful*, *sadly*), and hedges (*kind of*, *maybe*, *something like*), as well as other adverbial, adjectival, verbal, and modal markers of stance. The seed lexicon is augmented with markers from the Switchboard Dialog Act Corpus characteristic of spoken language, as is sometimes used in online discourse, such as *oh yeah*, *nah*, or *wow* (Jurafsky et al. 1998). Recognizing that online discussions differ in genre from both written texts and spoken language, Pavalanthan et al. (2017) "translate" the stance marker lexicon to the Reddit domain using computational techniques based on distributional statistics, drawing upon prior work on the expansion of sentiment lexicons. Specifically, they train word embeddings on a Reddit corpus using Wang2Vec (Ling et al. 2015), a structured skip-gram model, and then add tokens with a cosine similarity of at least 0.75 to the lexicon. This is done with the aim of adding "netspeak" variations of the stance markers to the lexicon. Then, they perform a multi-dimensional analysis (Biber 1992) to the distributional statistics of stance markers across subreddit communities, to isolate the main axes of variation across the stance markers.

These stance markers are not universal – they are latent patterns of co-occurrence between stance markers and subreddits in the corpus, and can only be delimited as such. For example, one of the stance dimensions identified within the corpus is characterized on one "extreme" by the relationship between discursive practices found on subreddits r/philosophy, r/history, r/science with stance markers *beautifully*, *pleased*, *thanks*, *spectacular*, and *delightful*, and on the other extreme by the relationship between r/pcmasterrace, r/leagueoflegends, r/gaming, and stance markers *just*, *even*, *all*, *no*, *so*. The stance dimensions do not necessarily manifest in semantically-meaningful or consistent ways, and do not generalize to other corpora. What they do show is that coherent groupings of stance markers can and do emerge across different online speech communities.

In later work, three different stance dimensions are predefined: *affect*, *alignment*, and *investment* (Kiesling et al. 2018). Instead of attempting to surface salient stance dimensions empirically and characterize them post hoc, the authors delineate these stance dimensions in theoretical terms. *Affect* is defined as the polarity or quality of the stance to the stance focus, *investment* is defined as a measure of epistemic stance, or how strongly invested in the topic the speaker is, and *alignment* is defined as a measure of the degree to which a speaker/writer aligns with their interlocutor(s), real or imagined. After manually annotating 1,265 utterances from 68 Reddit threads, Kiesling et al. (2018) train a classifier on the data to surface textual features that are predictive of high and low levels of each stance dimension. The reported tokens for each dimension are shown in 2.1.

Affect		Investment		Alignment	
High	Low	High	Low	High	Low
thank	please	!	little	thank	evidence
!	worse	tell	limit	limit	wrong
sing	everyone	hope	ink	other	able
noise	nothing	better	maybe	!	not
stop	entire	never	may	absolutely	opinion
friends	into	stick	wouldn’t	thanks	worse
good	burn	parents	everyone	now	mom
fiber	no	kept	know	so	be
kindle	password	.	wants	point	has
love	effectively	carefully	actual	some	well

Table 2.1: Top predictive tokens for the affect, investment, and alignment stance dimensions (Kiesling et al. 2018)

Although this approach benefits from the interpretability of the stance dimensions, as they are predefined and grounded in sociolinguistic theory, the predictive keywords are not robust enough to be used as features for downstream stance detection. Moreover, in order to generate predictive keywords for a custom corpus, a large sample must be manually annotated for labels across all stance dimensions. As Kiesling et al. (2018) detail, this can be a complex, laborious process, especally since the annotations often require highly subjective decision-making and the consideration of entire threads, rather than single utterances. Addressing this level of nuance in stance annotation remains an important open problem in the field.

2.2 Stance-taking and evidence-making

This thesis is interested in stance as it materializes in conversations mediating the legitimacy of evidence in the process of *becoming*. So, although there isn't necessarily a computationally-tractable way to operationalize stance (yet), it may still be possible to analyze how the authority of evidence is constructed and negotiated at-scale through the lens of a single stance dimension: epistemic stance, or evidentiality. By narrowing the scope of what we mean by stance, we seek to identify a set of linguistic markers with more concentrated explanatory power.

2.2.1 Defining epistemic stance and evidentiality

The relationship between epistemic stance and evidentiality is somewhat convoluted and ambiguous. Evidentiality has traditionally been understood as the aspects of an utterance that refer to the source of knowledge and the type of evidence that a speaker has for making a claim or assertion (Willett 1988). In approximately one-quarter of the world's languages, evidentiality is encoded in the grammar itself, and some studies consider only expressions belonging to an obligatory grammatical category to be evidential (Aikhenvald 2004). In languages like English, where evidentiality is not grammaticalized, a variety of linguistic resources may optionally deployed as "evidentials," including lexical, constructional, and interactional forms (Chafe and Nichols 1986; Fox 2001; Clift 2006; Sidnell 2012). Even with a more flexible definition of evidentiality, some scholars subsume evidentiality under epistemicity. Linguist Elise Kärkkäinen (2003) offers a few examples: Palmer (1986) explicitly includes evidentials under epistemic modality, arguing that evidentiality is just one way of coding commitment or lack of commitment towards the truth of the proposition being expressed. Biber et al. (1999) take a similar view, including the source of knowledge, or the perspective from which the information is given, as a category of epistemic stance marking. However, other scholars have written about evidentiality as the superordinate category (Chafe and Nichols 1986; Biber and Finegan 1989). As Kärkkäinen points out, "where evidentiality fits in with epistemicity and which one is considered the superordinate category varies from one researcher to the next," and is almost just a "matter of terminological convenience."

Mushin (2001) offers a helpful perspective on how the concepts are intertwined, describing *epistemological stance* as the way speakers deploy evidential markers to convey assessments of knowledge that are then absorbed into the interactional context. In other words, evidentials never just convey the source of knowledge or information – they always say something more (San Roque 2019). Direct evidential forms may be mobilized to "assert certainty or imply intimacy" with other participants in the conversation (San Roque, Floyd, and Norcliffe 2018); reportative forms may be used to index doubt, as the speaker distances themselves from the information and shifts responsibility to another interlocutor, whether real or imagined (Aikhenvald 2004); sensory or inferential forms may also be used as prosocial behavior (Mithun 1986). This suggests that evidential marking is "responsive to and constructive of the relationship between speaker and recipient(s)" (Fox 2001), and that evidentiality is itself interactional and dialogic.

2.2.2 Markers of epistemic stance and evidentiality

As discussed, there are a considerable number of linguistic resources that may be invoked for stance-taking, but the articulation of epistemic stance in conversational American English has been shown to be quite regular and routinized (Kärkkäinen 2003). In the Santa Barbara Corpus of Spoken American English, epistemic phrases such as *I think* and epistemic adverbs such as *probably* are the most frequently-used, and Kärkkäinen finds that epistemic stance (at least within the corpus) tends to be expressed by a relatively limited set of tokens, highlighting nine epistemic markers in particular: *I think*, *s/he said*, *I don't know*, *maybe*, *I said*, *I don't know* + compl., *I guess*, *I thought*, and *probably.*

Kärkkäinen observes that the range of semantic meanings expressed by epistemic markers is fairly wide, and clusters them into seven different groups: *reliability*, *belief*, *hearsay evidence*, *mental construct*, *deduction*, *induction*, and *sensory evidence.* This system is based on Chafe's description of the two dimensions towards knowledge: the reliability of the information, and the mode of knowing. Along the first dimension, epistemic markers (mostly epistemic adverbs) serve as an indicator on a continuum of reliability, from "very reliable" to "unreliable" (cf. *undoubtedly*, *surely*, *maybe*, *probably*, *might*, *may*). Along the second dimension, epistemic markers (mostly epistemic phrases) indicate where the knowledge is coming from, or the way in which knowledge is acquired (Chafe and Nichols 1986). Importantly, *reliability* is the only cluster where it is suggested that the epistemic marker itself determines the polarity of the stance taken. Is there, then, a consistent scale of epistemicity that could be used to characterize evidential advebrs in terms of the polarity of the stance?

Quirk et al. (1985) describe a system for assigning polarity to epistemic adverbs by classifying them as expressions of conviction and doubt, in a discussion of content disjuncts that modulate the degree to which the speaker believes that what is being said is true. These groupings are shown in 2.2.

Disjuncts expressing conviction	Disjuncts expressing doubt
Admittedly	Allegedly
Assuredly	Arguably
Avowedly	Apparently
Certainly	Conceivably
Decidedly	Doubtless
Definitely	Likely
Incontestably	Maybe
Incontrovertibly	Most likely
Indeed	Perhaps
Indisputably	Possibly
Indubitably	Presumably
Surely	Purportedly
Unarguably	Quite likely
Undeniably	Reportedly
Undoubtedly	Reputedly
Unquestionably	Seemingly
Clearly	Supposedly
Evidently	Very likely
Manifestly	
Obviously	
Patently	
Plainly	

Table 2.2: Disjuncts expressing conviction and doubt (Quirk et al. 1985)

Interestingly, Quirk et al. separately define a third grouping representing expressions that modulate the degree to which the speaker references the “reality” or lack of “reality” in what is said. While conviction and doubt might lie on a single axis of adverbial inflection, the third category requires further subdivision before polarity can be assigned. Adverbs like *actually*, *really*, and *factually* are considered to be assertions of the reality of what is said, while *only apparently*, *formally*, *hypothetically*, *ideally*, *nominally*, *officially*, *ostensibly*, *outwardly*, *superficially*, *technically*, and *theoretically* are listed as expressions that contrast with reality. *Basically*, *essentially*, and *fundamentally* are further differentiated as adverbs that claim that what is being said is true or real in principle. As becomes quickly apparent, defining a single illocutionary charge, or sign, for each epistemic adverb is not so simple. Even for just evidential markers, the polarity of the stance is refracted onto multiple dimensions, ranging from reliability to certainty to reality.

2.3 demonstrates the variation in how the sign is defined for some of the most discussed epistemic adverbs across various studies. While some of the systems seem to have coherent and consistent labels – Wierzbicka’s allocation of “confident” (2006) roughly maps to Quirk et al.’s “conviction” (1985) and Huddleston and Pullum’s “strong certainty” (2002) – there are still significant differences in coverage and sometimes, in definition. For example, Huddleston and Pullum treat *evidently* as a "low certainty" adverb, while Wierzbicka and Quirk et al. mark it as a sign of confidence and conviction. Biber et al. (1999) also categorizes some adverbs with the "Source of Knowledge" (SOK) label, which is not represented in the other systems.

Because evidentiality is interactional, and has more diverse manifestations than is apparent from the composition of the utterance alone, this thesis does not attempt to assign polarity to evidential adverbs to measure the directionality of the stance being taken (despite the substantive appeal). Instead, with evidential adverbs as the point of departure, we turn our attention empirically to how they are mobilized in online discourse. Although they are not deterministic indicators of stance-taking per se, we find that it is still possible to trace how certain evidential adverbs in social media discourse significantly impact the contours of the social and interactional context they

Adverb	Wierzbicka	Quirk	Biber	Huddleston
Allegedly	Hearsay	Doubt		
Apparently	Hearsay	Doubt	SOK	Low certainty
Arguably		Doubt		Low certainty
Certainly		Conviction	Certainty	Strong certainty
Clearly	Confident	Conviction		Strong certainty
Conceivably	Nonconfident	Doubt		Lowest certainty
Evidently	Confident	Conviction	SOK	Low certainty
Likely		Doubt	Likelihood	Low certainty
Obviously	Confident	Conviction	Likelihood	Strong certainty
Possibly	Nonconfident	Doubt		Lowest certainty
Presumably		Doubt		Low certainty
Probably			Likelihood	Low certainty
Reportedly	Hearsay	Doubt		
Seemingly		Doubt		Low certainty
Supposedly	Hearsay	Doubt		
Undoubtedly		Conviction	Certainty	Strong certainty
Unquestionably		Conviction		Strong certainty

Table 2.3: Variation in descriptive categorizations of epistemic adverbs

linguistically co-construct. In the following chapters, we discuss how the practice of evidentiality (and the process of *becoming evidence* more broadly) manifests in sites of controversy, shaping the landscape of possibility for participation in online debates where the negotiation of data-as-evidence is essential and even, at times, existential.

Chapter 3

The truth is out there

This chapter looks at how evidentiality materializes on social media through a close examination of a specific speech community on Twitter, identified by the use of the hashtag #ufotwitter. Since 1947, reports of unidentified flying objects and alien encounters have captured the public's curiosity, inspiring amateur research (also known as ufology), extra-terrestrial contact support networks, government investigations, bestselling books, and sensational news coverage (Eghigian 2017; Pasulka 2019). Now, more than seven decades later, much of this conversation continues online. On Twitter, Reddit, and other internet forums, entire communities have formed to discuss government reports, new sightings, and conspiracy theories surrounding what they consider to be one of the world's most important secrets. This seemingly boundless stream of debate and speculation hinges on the belief that, as *The X Files* famously suggested, the truth is out there – that is, the truth is not yet here. In #ufotwitter we find a rich site of evidential discourse suggesting that truth, in fact, is not something to be found and disclosed, but rather negotiated and mediated over time. Using a mixed-methods approach, we identify patterns of linguistic variation in this discursive community that modulate how data is circulated and taken up as evidence online.

3.1 Ufology explained

Ufology, broadly, is the study of unidentified flying objects (UFOs), also known as unidentified aerial phenomenon (UAPs). Historians of American ufology usually trace it back to 1947, when US Army Air Forces officers mistakenly identified debris from a weather balloon crash near Roswell, New Mexico as a "flying disc." Although they quickly retracted their statement, public interest in the incident remained. Ufologists entertained other possibilities to explain it, speculating that private military contractors and clandestine, para-governmental organizations had covered up the evidence to conceal its extraterrestrial origin (Lewis-Kraus 2021).

Historically, academic researchers have rejected ufology as a discipline as wrongheaded, irrational, and dangerous. The presence of conspiracy theory and paranormal belief within ufology has reinforced the general impression that the movement is shrouded in paranoia and mysticism. This, in turn, has contributed to its marginalization as a subject unworthy of serious professional consideration (Eghigian 2017; Pasulka 2019). The government has also routinely responded to UFOs with expressions of indifference and dismissal (Lewis-Kraus 2021). Ufologists have always been well aware of their illegitimate status within scientific and public policy circles – leading them to seek alternative communicative strategies, bypassing institutional authorities by either speaking directly to the general public via mass media, or founding parallel sites of knowledge production with their own methodologies and peer reviews. This has led to a deep culture of mutual mistrust between ufologists and institutional authorities (Eghigian 2017).

The suspicion that the government keeps knowledge of or about UFOs from its citizens is not entirely unfounded. In 2017, the *New York Times* ran a front-page story revealing that the Pentagon had been running a surreptitious UFO program for ten years – not the first of its kind. After the Roswell incident in 1947, reported sightings of unexpected and unfamiliar things in the skies apparently became too profuse for the Air Force to ignore. Lieutenant General Nathan F. Twining wrote what is now well-known in ufologists' circles as the "Twining Memo," asserting that "the phenomenon reported is something real and not visionary or fictitious" and alluding to concerns that a great power rival such as the Soviet Union could be behind it all. The government launched Project Sign to investigate. This initiative, and its successor, Project Blue Book, culminated in new protocols for recording reported cases of UFO sightings and encounters as standardized data, and the systematic review of approximately twelve thousand of these cases – seven hundred and one of them of which were unresolved. Subsequent programs such as the Advanced Aerospace Threat Identification Program, or AATIP, focused on the national security implications of military UAP encounters. Given the historical setting of the sightings, the secrecy with which these programs were enshrouded is perhaps unsurprising. Post-World War II and Cold War anxieties made it such that reports of UFOs were quickly folded into the enterprise of intelligence analysis by governments, and the public – including most civilian scientists – were to be informed strictly on a need-to-know basis ((Eghigian 2017; Lewis-Kraus 2021).

Upon this backdrop of state secrecy, disclosure – the release of classified information – has become incredibly salient (Lepselter 2016). For ufologists, disclosure signifies multiple things. It is not just the publication of precious and previously inaccessible evidence, or the government's admission to a persistent interest in something they had consistently snubbed. In a setting where, as Espírito Santo and Vergara (2020) put it, evidence is also the constructor of possible worlds, disclosure becomes the creation of a new frontier of perception and possibility.

In the months leading up to the release of the Pentagon UFO Report by the Unidentified Aerial Phenomena Task Force (UAPTF) in June 2021, many wondered if the first real instance of disclosure was finally at hand. In a sense, those that were expecting a singular, discrete event – the disclosure of truth – were severely let down. The highly anticipated report, which assessed 144 cases from 2004 to 2021, was cautious in its pronouncements, finding that only a handful of these incidents demonstrated significant and inexplicable movement patterns of flight characteristics (indicative of the presence of advanced technology), and that "limited data leaves most UAP unexplained" (Director of National Intelligence 2021). However, the report also called for more resources, more investment, and above all, more data.

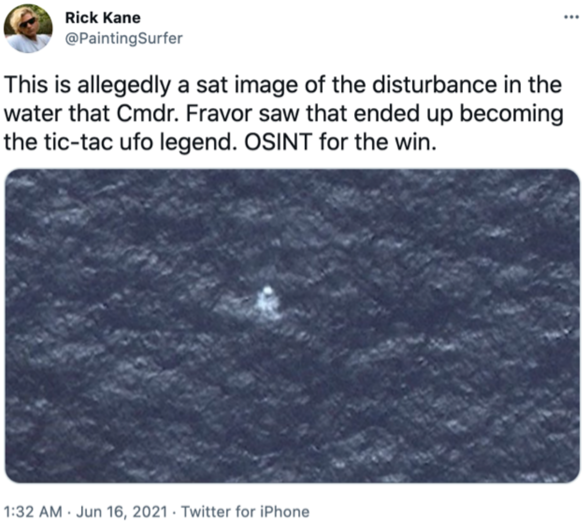
Given the prominent position "data" is given the report, it is worth noting that its definition of what constitutes legitimate data is quite vague. The UAPTF does specify that the dataset only consists of US Government reporting of incidents, and its call for dedicating more optical and radiofrequency sensors to data collection (to capture the relative size, shape, and structure of UAPs, as well as more accurate velocity and range information) suggests that it may privilege data from these sources over others. Yet the extent to which it might do so is unclear. Although the directive given to the UAPTF in Senate Report 116-233 includes a request for the detailed analysis of unidentified phenomenon data collected by “geospatial intelligence, signals intelligence, human intelligence, and measurement and signatures intelligence,” the degree to which each of these sources of information appears in the UAPTF’s dataset is not revealed. Despite these ambiguities, data seems to be generally accepted as the means through which clarity and consensus will be achieved, and doubt eliminated, with government officials even citing a willingness to “go wherever the data takes us” (Kube and Edelman 2021). This shift, however tiny, in ufology’s frontiers of perception and possibility reminds us that disclosure is just one point in the continuum of evidentiality, where the boundaries of what is conceivable and inconceivable are constantly negotiated anew (Agrama 2021; Espírito Santo and Vergara 2020). In the following sections, we discuss how these disclosures are transacted in online media, and consider what a statistical analysis of evidentiality in ufologists’ discourse can (and cannot) reveal.

3.2 Ufology online

Without a standard-setting institution for how data should be accepted as evidence, or even for what counts as data, ufology has experienced a “proliferation of local modes of knowledge-making, each of which has its own unique protocols” (Espírito Santo and Vergara 2020). Moreover, without academic or political legitimacy, ufologists have often sought to express their views by speaking directly to the general public via mass media (Eghigian 2017). These features of communicative practice tend to collide to spectacular effect in the fast-paced and volatile information ecologies found in social media. In this section, we look into discursive communities formed by ufologists on Twitter for further insight into how evidentiality is practiced online.

While most academic scientific discourse on Twitter is focused on distributing scientific information to a public audience (Mandavilli 2011; Runge et al. 2013; Choo et al. 2015), ufologists' interactions on the platform almost function as an intensified peer review. That is, in addition to circulating and contesting evidence, they are constructing it, too. Consider the tweet referenced in 3-1, in which Twitter user @PaintingSurfer writes: "This is allegedly a [satellite] image of the disturbance in the water that Cmdr. Fravor saw that ended up becoming the tic-tac ufo legend. OSINT for the win" (PaintingSurfer). This is a reference to the "Nimitz encounter," a sighting of "a white oval object that resembled a large Tic Tac... about forty feet long, with no wings or obvious flight surfaces and no visible means of propulsion" from 2004 (Lewis-Kraus 2021). The sighting was reported in the midst of the Nimitz Carrier Strike Group's training operations by Commander David Fravor, and was corroborated by two other pilots. Yet another pilot was dispatched to attempt to record the object after Fravor's sighting, and the resulting video – one minute and sixteen seconds of blurry, monochromatic action – has been the subject of extensive media coverage since 2017.

@PaintingSurfer's tweet is a microcosm of the contradictory tensions that surround this particular controversy. The use of the epistemic adverb *allegedly*, for example, amplifies uncertainty regarding the link between the satellite image attached to the tweet and the Nimitz Encounter itself. Furthermore, the use of the word *legend* implies some level of doubt, reinforced by the preceding phrase *ended up becoming*, which suggests reflexively that the legend was socially constructed – and perhaps something more than, or different from, the event that was witnessed. However, @PaintingSurfer also comments: "OSINT for the win." This is a reference to open-source intelligence, or the practice of collecting information from published or otherwise publicly available sources. Taken as a whole, @PaintingSurfer's epistemic stance on the matter is

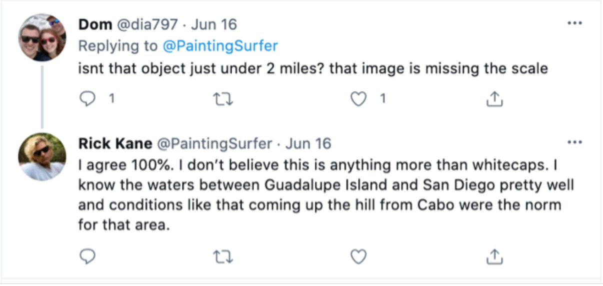


Rick Kane
@PaintingSurfer

This is allegedly a sat image of the disturbance in the water that Cmdr. Fravor saw that ended up becoming the tic-tac ufo legend. OSINT for the win.

1:32 AM · Jun 16, 2021 · Twitter for iPhone

Figure 3-1: Tweet of a satellite image linked to the Nimitz encounter



Dom @dia797 · Jun 16
Replying to @PaintingSurfer
isnt that object just under 2 miles? that image is missing the scale

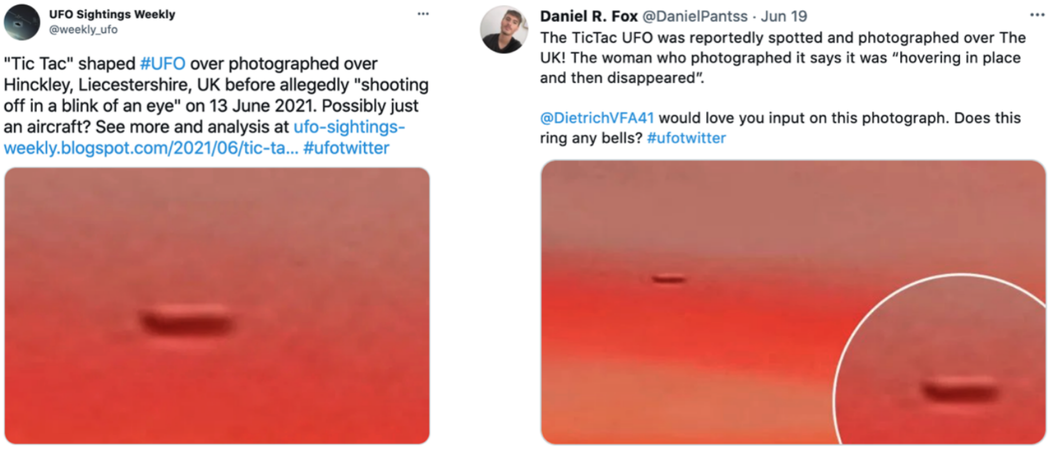
Rick Kane @PaintingSurfer · Jun 16
I agree 100%. I don't believe this is anything more than whitecaps. I know the waters between Guadalupe Island and San Diego pretty well and conditions like that coming up the hill from Cabo were the norm for that area.

Figure 3-2: A response to @PaintingSurfer's original tweet

unclear.

Of course, the act of publishing this on Twitter invites further scrutiny from other amateur researchers, enthusiasts, and skeptics. An example of this is shown in 3-2, a direct reply to @PaintingSurfer's original tweet calling into question the lack of a scale in the image. @PaintingSurfer's response to this sheds further light on his personal stance on the matter: "I don't believe this is anything more than whitecaps." He then goes on to employ experiential evidence to justify disbelief in the image.

How do we make sense of @PaintingSurfer's original tweet given this contradictory context? Does the doubt implicated by the phrase *this is allegedly a sat image* outweigh the affirmation of OSINT's role in this ecosystem of evidential production? Is he applauding OSINT for simply surfacing the image for further consideration, or is he actively promoting its evidential authority? Was he suspicious of the image from



UFO Sightings Weekly
@weekly_ufo

"Tic Tac" shaped #UFO over photographed over Hinckley, Liecestershire, UK before allegedly "shooting off in a blink of an eye" on 13 June 2021. Possibly just an aircraft? See more and analysis at ufo-sightings-weekly.blogspot.com/2021/06/tic-ta... #ufotwitter

Daniel R. Fox @DanielPantss · Jun 19
The TicTac UFO was reportedly spotted and photographed over The UK! The woman who photographed it says it was "hovering in place and then disappeared".

@DietrichVFA41 would love you input on this photograph. Does this ring any bells? #ufotwitter

Figure 3-3: Two tweets about a Tac Tac UFO sighting in the UK

the beginning, or did the interjection from @dia797 convince him otherwise? One possible interpretation (of many) is that @PaintingSurfer personally doubted from the outset that this particular instance of satellite imagery was legitimate evidence, but opted to circulate it, and shift the focus of his initial tweet on the merits of OSINT instead. Regardless of whether or not this was the case, the juxtaposition of these positions once again puts into sharp relief how nuanced and circumstantial the articulation of stance can be.

Sometimes, ufologists on Twitter will explicitly call upon each other in an effort to invoke the peer review process. This is done in a couple different ways, as shown in 3-3. Some mobilize hashtags such as #ufotwitter to make sure that the tweet reaches the proper audience. Others will directly tag members of the community with authority in an attempt to capture their attention and get feedback on new data.

Twitter users started to gather around the hashtag #ufotwitter in December 2017, likely in response to the *New York Times* article that broke the news of the Pentagon's UFO program (Cooper, Blumenthal, and Kean 2017). Although #ufotwitter has not received much scholarly or popular media attention, insight into the community can be found in reflexive discussions led by members on podcasts, blog posts, Youtube, and #ufotwitter itself. One ufologist, AP Strange, describes it as follows on the *Chaos and Shadow* podcast (KyleParanormal and Pagan 2021):

> For the last couple of years I've tactically ignored a lot of the discussions

> that #ufotwitter has. They just really bore the hell out of me, to be honest. It's a lot of material, nuts and boltsy, government documents, and hoping for some kind of utopian disclosure movement, some kind of moment where this is gonna happen and the reality of something, I dunno – not UFOs, not UAPs, just the disclosure that the US government has some kind of alien technology – [comes out.] I don't even know anymore... Disclosure is always going to be the football that Lucy van Pelt is holding up, and #ufotwitter is ever the Charlie Brown running up to kick it.

A Youtube video by Engaging the Phenomenon (2020) offers a more approbatory perspective:

> It's a diverse group of individuals who are both like minded and... quite different, but we're all at the same time (regardless of our own research and opinions) sharing information and engaging in conversations. Some #ufotwitter participants are breaking stories, putting together new information. And then right behind them you have other participants of ufo twitter that take those stories and leads and follow up and put together more data and information, and we have a pool of information going on through this activity. UFO twitter is very fast paced. It's also very limited in the text you can use, so you have to be fairly concise and to the point in the short limit of text you can actually use. Believe it or not, that actually works very well for some reason, in my opinion.

In the video, in which members are interviewed and asked to describe the community in their own words, #ufotwitter emerges almost as an institution in and of itself. According to Twitter user @SteUFOnotCGULLS, #ufotwitter has "grown and grown, it's gone worldwide now. *It's like a think tank*" (all emphasis mine). @Deepneuron notes that "the beautiful thing about ufotwitter is that it's a group of people who are constantly exchanging ideas and looking for the truth as what people will see as *the world's most important secret.*" @akam1129 emphasizes a need for the community to be welcoming: "We have to help all people who walk onto the subject –

the political subject, the academic subject, scientific... and everybody is concerned, it's very important. *It's a great communion*, to be here all together." @mikeb8637 adds that "When I think about what #ufotwitter is to me, I envision it as *a vehicle for participation* in the topic of the phenomenon, and instead of creating barriers for entry into this topic it removes them, for better or for worse."

These characterizations of #ufotwitter invoke political, academic, and even religious language (cf. Pasulka 2019), perhaps reflecting a latent desire within the community to break free of the institutional isolation that has been imposed upon it. In a way, this has already been done on social media. By bringing evidentiary protocols and perspectives online, #ufotwitter has become that institution for those seeking it.

3.3 Evidentiality on ufotwitter

In the subsequent sections, we present a corpus-based analysis of the linguistic dimensions of evidentiality on #ufotwitter. We introduce a novel dataset of tweets using the hashtag #ufotwitter from 2019-2021, and document work done to answer two initial research questions regarding the corpus:

> **RQ1.** Do different subcommunities within #ufotwitter mobilize language differently when talking about evidence?
> **RQ2.** How does evidential language mediate interaction and participation on #ufotwitter?

We find that #ufotwitter is small and niche enough that many of the most active members will engage with each others' content across ideological boundaries, making the disaggregation of the corpus into discrete subcommunities difficult using common network analysis algorithms or available Twitter metadata. We also find that coding tweets with sociologically-meaningful notions of interaction is difficult to do at-scale. However, in these efforts to render the research questions tractable, we discover new paths and insights into the ways in which the language of evidentiality shapes public perceptions of truth and reality.

3.3.1 Data collection

Using the Academic Research search endpoint of the Twitter v2 API, we collect a corpus of more than 3 million tweets using the keyword “ufo.” However, we find that this corpus is very noisy, including many tweets discussing musical bands, movies, games, and other media. In order to focus our analysis on talk that is (mostly) about evidence, we narrow our search query to a single self-selecting speech community and retrieve only tweets that use the hashtag #ufotwitter. This strategy yields a corpus of 230,000 tweets.

The search parameters used are shown in 3.1. The query makes use of the implicit AND operator described in the Twitter query-building guidelines and filters out tweets with keywords "bts" and "fortnite" (which, by manual inspection, were determined to be significant contributors of noise to the dataset). The query is also designed to filter out retweets with the "-is:retweet" clause, resulting in the collection of only original tweets, quote tweets, and replies. Because this is a full-archive search, returning many more results that can be stored in a single API response, pagination is used to fetch the data in a series of “pages.” The tweet data, user metadata, and media metadata are disaggregated from the response and serialized and saved separately.

Search parameter	Value
`query`	#ufotwitter -bts -fortnite -is:retweet
`tweet.fields`	id, text, author_id, conversation_id, entities, in_reply_to_user_id, referenced_tweets, attachments, created_at, public_metrics
`expansions`	attachments.media_keys, author_id, in_reply_to_user_id, referenced_tweets.id.author_id
`media.fields`	media_key, type, url
`start_time`	2006-03-21T00:00:00Z
`user.fields`	id, location, name, public_metrics, url, username, verified

Table 3.1: Search parameters for full-archive search using the Twitter v2 API

3.3.2 Evidential markers

For both research questions, we account for evidential language with the frequency of specific evidential markers. Following the literature discussed in Chapter 2, we focus on how twelve evidential markers (listed in 3.2) are deployed within speech on #ufotwitter.

Evidential markers
Allegedly*
Apparently*
Arguably
Basically
Certainly
Clearly
Evidently
Possibly
Presumably
Reportedly*
Seemingly
Supposedly*

Table 3.2: Evidential markers of interest (*reportative adverbs)

Within this set, we also pay special attention to *allegedly*, *supposedly*, *reportedly*, and *apparently*, which are reportative adverbs and more predictable in function (compared to, for example, multivalent adverbs such as basically, which doubly functions as a common discourse marker). According to Martin and White (2005), reportative adverbs function as resources for attribution – that is, they are used to "attribute the proposition to some external source." This function situates them within the realm of evidentiality (Rozumko 2019), and makes them of particular interest to our study.

3.4 Variation in evidential inflections on ufotwitter

To investigate **RQ1**, we decompose the question into two parts: (1) can we segment the corpus into discrete, meaningful subcommunities; and (2) can we compare the statistical distribution of the markers listed in Section 3.3.2 to gain insight into how evidential inflections are differently mobilized between groups? Are tweets mobilizing

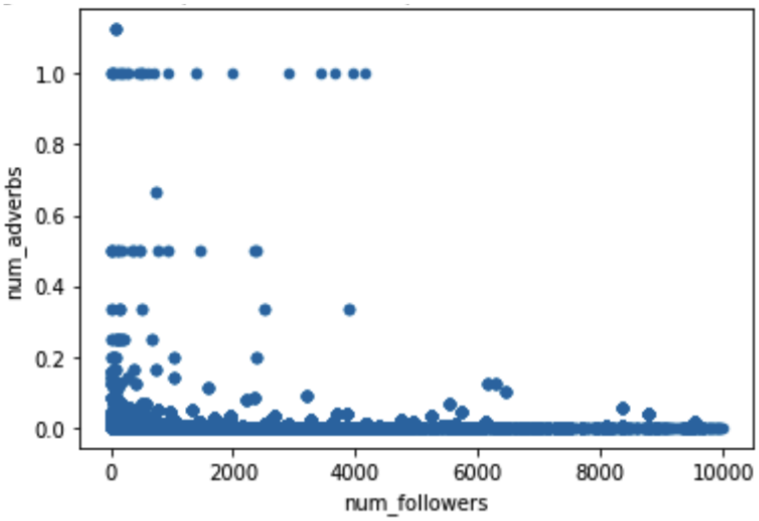


Figure 3-4: Social status and reportative adverb deployment in the #ufotwitter corpus

evidential language differently depending on characteristics of the user or the text itself, and to what sociolinguistic effect? In this section we discuss three methods of segmenting the corpus: with follower counts from user metadata as a proxy for social status, with user interaction graphs, and with topic modeling. For each, we discuss motivation, implementation, and preliminary results.

3.4.1 User metadata

Do members of #ufotwitter with many followers employ evidential language at different rates than those with a smaller number of followers? Here, we use follower count metadata for the authors represented in the corpus in order to estimate "social status" within the dataset. We harmonize the tweet dataset with the user dataset saved from the data collection phase, and compute the number of reportative adverb tokens present in each tweet. The relationship between these two measures is plotted in 3-4. We find that they are roughly negatively correlated, suggesting that users with higher social status use reportatives at a lower frequency. However, we determine that the relationship is too tenuous to draw conclusions or justify further quantitative analysis.

3.4.2 User interactions

In this section, we attempt to surface latent subcommunities in the user network structure before comparing their employment of evidential language on #ufotwitter. By building a user graph based on interactions within the dataset (who retweets/replies to/mentions whom), we test an underlying assumption that Twitter users who share ideological predilections or other (offline) social characteristics tend to talk to each other more than they talk to people outside of their affiliated groups. Ultimately, we find that this assumption does not hold for our corpus, but we document our efforts nonetheless.

To begin, we construct a user graph using the user metadata from the corpus, where each node corresponds to a user who appeared in the dataset. This is an unweighted, undirected graph; edges are drawn between two nodes (users) if one has retweeted, quoted, replied to, or mentioned the other. This produces a graph with 36,069 nodes and 127,592 edges, with an average degree of 7.07. We then run community detection on the graph using the Louvain method (Blondel et al. 2008), an algorithm that optimizes the modularity of a network. The "betweenness" measure of modularity (Q) proposed by Newman and Girvan (2004) is defined as follows:

$$Q = \sum_i (e_{ii}\text{-}a_i^2)$$

Here, e_{ij} is the fraction of all edges in the network that connect vertices in partition i to those in partition j, and a_i is the fraction of edges that connect to vertices in partition i:

$$a_i = \sum_j e_{ij}$$

The modularity Q is effectively the fraction of edges in the network that connect vertices of the same type (that is, within-community edges) minus the expected value of the same quantity in a network with the same community divisions but random connections between the vertices. A lower value indicates weak community structure (no better than random) while values approaching 1 (the maximum) indicate a strong

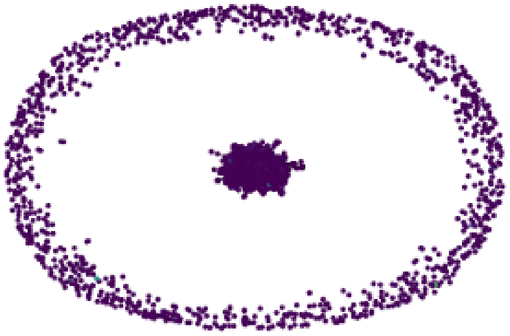


Figure 3-5: Community detection on the base user graph

community structure. The Louvain algorithm greedily optimizes modularity by progressing iteratively with a two-step pass: in the first step, modularity is maximized by allowing only local changes of communities; then, the communities are aggregated to build a new network of communities.

For the base graph, the Louvain method does not yield a discernable grouping amongst the users, as shown in 3-5.

Given that 58% of the nodes in the base graph are of degree 1, meaning that they are connected to just one other node, we compute subgraphs constrained by degree (so that subgraph G_k contains only nodes with degree k or higher, where k ranges from 2 to 10). This is done with the aim of running the Louvain method on denser, smaller graphs. However, constraining by degree only results in minimal improvements in community detection, whether $k = 2$ (3-6) or $k = 10$ (3-7). This can be seen not only visually and in the modularity scores themselves, but also by inspecting the allocation of users in the communities detected – known skeptic accounts like @MickWest are routinely grouped with known enthusiast accounts such as @uncertainvector or @AFSUnidentified.

Further inspection suggests that the small and niche character of the #ufotwitter community makes modularity an unsuitable metric for optimization in the user graph: as illustrated in 3-8, members of #ufotwitter from opposite ends of the ideological spectrum are relatively willing to engage with each other, perhaps moreso than in other sites of controversy on Twitter. As US Navy pilot fighter Alex Dietrich puts it, "no matter how much they're attacking each other [on #ufotwitter]... they all want

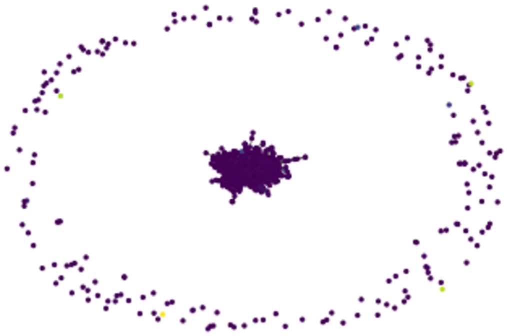


Figure 3-6: Community detection on a subgraph of nodes with degree 2 or higher

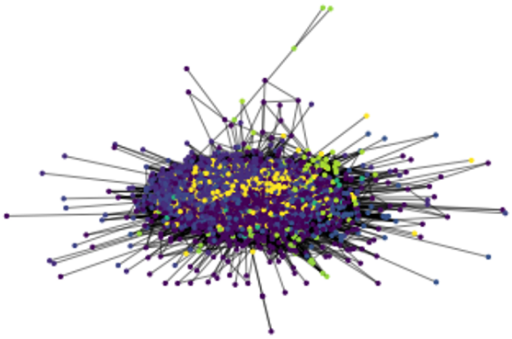
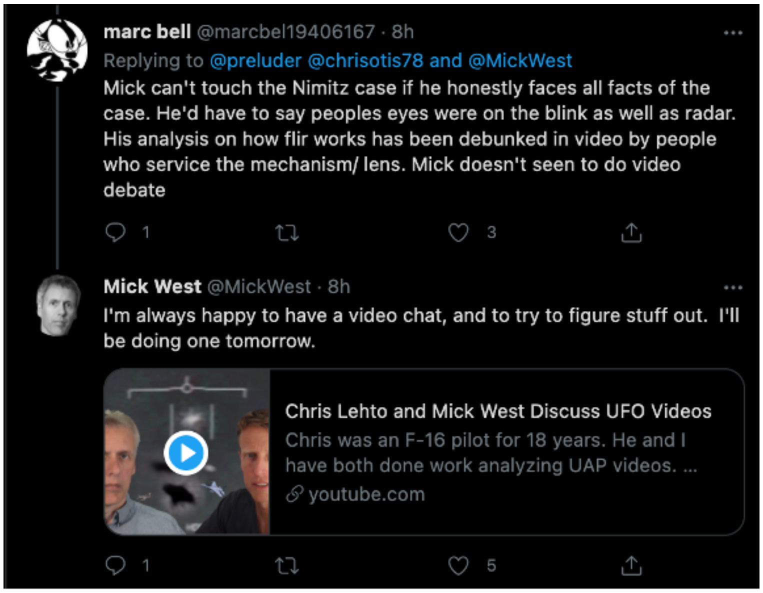


Figure 3-7: Community detection on a subgraph of nodes with degree 10 or higher

answers" (Dietrich).

3.4.3 Topic modeling

We also attempt to segment the corpus by looking for content-based clustering patterns in the text. We use Latent Dirichlet Allocation (LDA) with standard hyperparameters and $K = 3$ (K fixes the number of clusters) as the topic model, and fit the model to a lemmatized version of the corpus with nouns only, using the NLTK TweetTokenizer, WordNet lemmatizer, and POS tagger in the preprocessing pipeline. The top 5 most salient terms are *video*, *time*, *phenomenon*, *article*, and *news* for topic 1; *space*, *alien*, *science*, *disclosure*, and *spacex* for topic 2; and *report*, *ovnis*, *pentagon*, *book*, and *life* for topic 3 (Appendix A). This suggests that talk concerning observed and experiential evidence is allocated to topic 1, and that discourse about government reports and other authoritative sources of evidence is captured by topic 3. However, the associations are not strong enough to effectively segment the corpus using the



marc bell @marcbel19406167 · 8h

Replying to @preluder @chrisotis78 and @MickWest

Mick can't touch the Nimitz case if he honestly faces all facts of the case. He'd have to say peoples eyes were on the blink as well as radar. His analysis on how flir works has been debunked in video by people who service the mechanism/ lens. Mick doesn't seen to do video debate

Mick West @MickWest · 8h

I'm always happy to have a video chat, and to try to figure stuff out. I'll be doing one tomorrow.

Chris Lehto and Mick West Discuss UFO Videos

Chris was an F-16 pilot for 18 years. He and I have both done work analyzing UAP videos. ...

youtube.com

Figure 3-8: Mick West, a prominent skeptic and debunker, interacts directly with a #ufotwitter enthusiast

topic model alone. This is further corroborated by the lack of significant deviation in evidential adverbial usage across the topics, as shown in 3-9.

3.5 Variation in evidential production on ufotwitter

For **RQ2**, we endeavor to study how evidential language mediates interaction and participation using two methods. First, we look at variation in the *focus* of evidential modulation across the corpus, paying special attention to different participatory

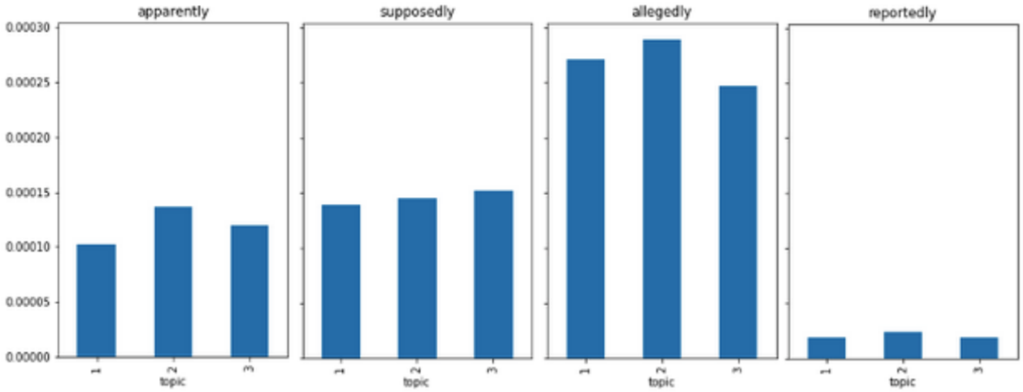


Figure 3-9: Frequency of reportative adverb deployment across #ufotwitter topics

aspects of evidentiality. Second, we look at variation in the rates of engagement associated with the use of evidential adverbs in the tweets.

3.5.1 A participation framework of evidential production

To break down how the evidential adverbs mobilized on #ufotwitter encode different participatory aspects of evidentiality, we introduce a participation framework of evidential production. This framework delineates how different focal subjects of evidential modulation can be identified when evidential adverbs are employed in speech. Drawing upon Goffman's participation model, an expansion of the traditional speaker-hearer dyad in interactive discourse, we define six entities that participate in the dialogic process of evidential production: the *phenomenon* itself, the *representation* of the phenomenon, the *author* or witness of the evidence at hand, the *animator* of the evidence, the *principal* (the party who is socially responsible for the evidence), and the *audience* for the evidence. Our model tries to isolate the subject of the stance being taken when an evidential adverb is invoked, and slot it into one of these categories. Using this granular sociolinguistic coding schema allows for a more nuanced investigation into the performative elements of becoming evidence. Rather than taking the one-dimensional view that something is worthy of being considered evidence on the basis of being true or false, real or fake, such a framework scrutinizes multiple representational and social aspects of data that must be rendered legible, credible, and then accepted before becoming evidence.

We manually annotate a dataset of 800 tweets from the #ufotwitter corpus, filtered by the presence of an evidential adverb, with the coding schema described above. Two rounds of co-annotation are done with two people independently coding small subsets of tweets (40 in the first round, 10 in the second round). The annotations are compared with a third party to resolve disagreements as we refine the coding framework and strategy. After intercoder reliability is achieved with a Krippendorff's alpha of 0.882, the remainder of the tweets are annotated on a single-coder basis.

3-10 shows the distribution of these codes across the annotated corpus. The evidential adverbs are found to modulate the *phenomenon* and *representation* most

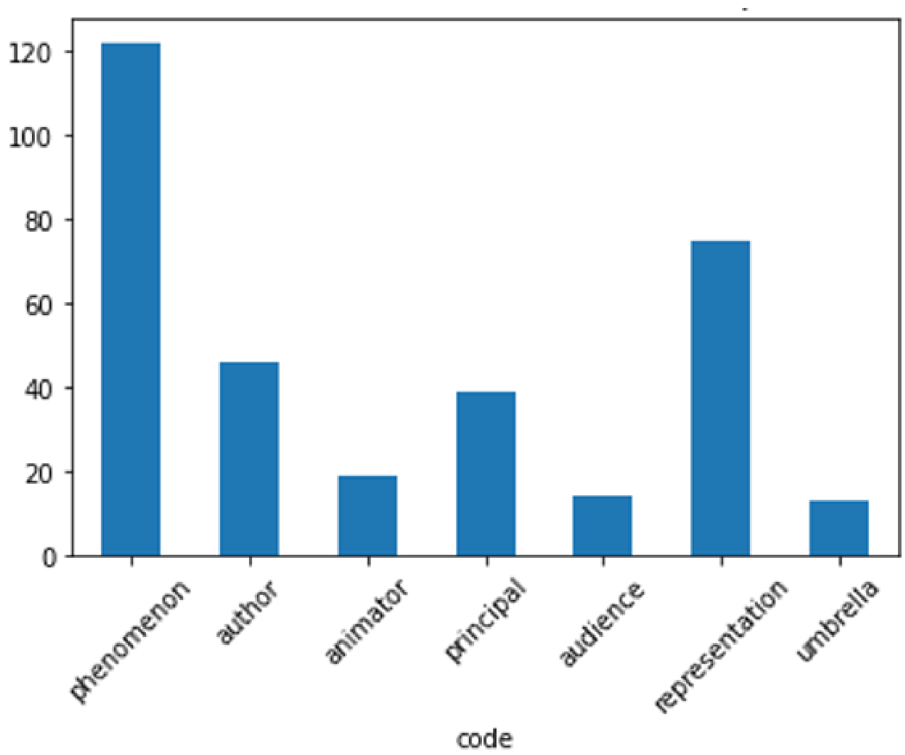


Figure 3-10: Distribution of participation codes in the annotated corpus

frequently in the dataset. Still, a nontrivial number of tweets focuses attention on other actors involved in producing or circulating evidence, including the original observer (*author*) and the subsequent reporter (*animator*). This distribution suggests that the language used on #ufotwitter to encode stance reflects a highly nuanced awareness, emergent on the syntactic level, of the multifaceted nature of evidence.

Naturally, the question of whether specific evidential adverbs are mobilized at different rates for certain participatory effects follows. Although there is some statistical variation, as shown in 3-11, no distinct patterns emerge at this sample size.

3.5.2 Engagement and evidential production

Finally, we consider how the language of evidentiality impacts engagement as measured by the affordances of interaction on Twitter: with replies, likes, retweets, and quotes. Are certain evidential adverbs mobilized to differential participatory effect?

In the 800-token annotated set described above, we find notably higher-than-average rates of engagement for tweets mobilizing a subset of the reportative adverbs.

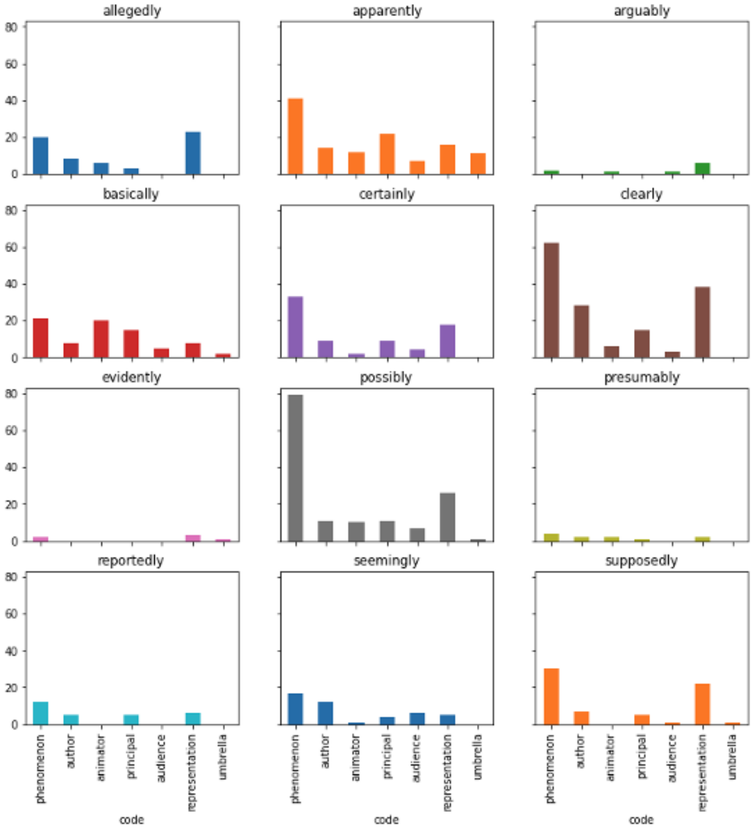


Figure 3-11: Distribution of participation codes disaggregated by evidential adverb

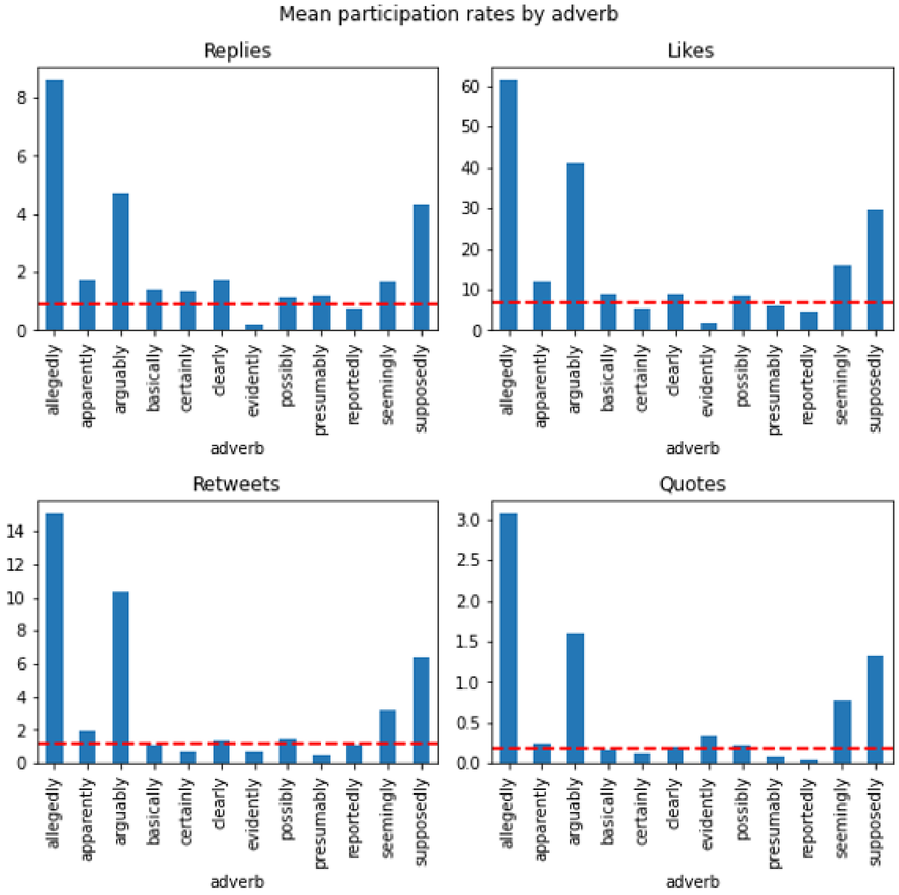


Figure 3-12: Mean participation rates by evidential adverb

We plot the mean participation rates for each evidential adverb against the mean participation rate across the entire #ufotwitter corpus (demarcated by the dotted red line) in 3-12.

Some adverbs, such as *allegedly* and *supposedly*, appear to have an outsized impact on levels of engagement with the evidence on #ufotwitter. These adverbs are particularly significant in the sociolinguistic context, not only because of their pragmatic function as *reportative adverbs*, but also for being indexical of the circulation of *hearsay*. What does it mean, then, that assertions of evidence that call conspicuous attention to weakness rather than strength tend to generate more engagement? In the next chapter, we consider the *hearsay effect* through various lenses, assessing questions of causality and generalizability across different domains of controversy.

Chapter 4

The hearsay effect

In the preceding chapters we traced evidentiary practice on social media in search of a computational instrument to measure stance, only to arrive at something we call the hearsay effect. This is the observation that certain evidentials, referred to herein as hearsay adverbs, draw disproportionately high levels of engagement (as indicated by metrics such as retweets and replies) when mobilized on #ufotwitter. This chapter examines the hearsay effect in greater depth by asking three interrelated sets of questions:

> (1) What does it mean, socially, when evidence is presented as hearsay in social interaction? What does a high level of online engagement signify in this context?
> (2) Is engagement responsive to participatory inflections in the language (the hearsay adverbs themselves) or is it reflective of other linguistic features that are not computationally-legible?
> (3) Can the hearsay effect be detected in multiple discursive communities? Is it the same across contexts, or is the hearsay effect more or less salient for specific kinds of discourse?

Using a mixed-methods approach, we show that the hearsay effect is variable across different interactional contexts and correlative with a particular genre of online evidentiality that is characteristic of open source intelligence. Although the hearsay effect is not itself a predictor of stance, it is useful as an index of an evidentiary genre that arises in the absence of epistemic authority.

4.1 On hearsay

Hearsay is defined epistemologically as indirect or second-hand testimony obtained from a source that is neither present nor accountable for what is being said (Bakhurst 2013; McDowell 1998). It can be analyzed as a communicative process on the triadic level, in which a speaker S conveys a proposition p to a recipient R where p has not been produced or designed directly by S, but is rather derived elsewhere (Martini 2017). In all the examples we analyze, the addition of a hearsay adverb overly signals that proposition p has been other-authored.

The expression of hearsay is an instance of stance-taking. Whether the epistemic validity of the proposition is accepted by the recipient determines the alignment between the speaker and the recipient. This process is related to epistemic properties of the speaker's utterance, as well as the recipient's own knowledge of the context.

Consider the following conversation between Sam and Rachel, where Sam says: "I heard Pepe got a video of a UFO hovering over the Green Building last night." The truth of the utterance – that Pepe captured a video of a UFO, or that there was a UFO hovering over the Green Building in the first place – is undetermined. Sam does not commit either which way, but ultimately leaves it up to Rachel to form a judgment of her own on the matter.

Rachel might respond to this in a number of different ways. She might accept the assertion at face value and leave it at that – that Sam said there is some chance that Pepe captured a video of a UFO last night. She might accept it as an expression of belief – that Sam believes that Pepe captured a video of a UFO last night. She might also accept it as knowledge – that Pepe did, in fact, capture a video of a UFO last night. This choice, and the degree to which Rachel commits to this choice, depends not only on the inflections of certainty or doubt interpreted from Sam's utterance, but also on her relationship to her interlocutor and the context (i.e., the Goffmanian participation framework) more broadly – does she trust Sam? Does she trust Pepe? Does she trust video evidence, or believe that UFOs exist (and can be captured on camera)?

This example once again shows us that the points of epistemic modulation in hearsay are multiple and variegated. It also demonstrates that the expression of hearsay is necessarily dialogic. Scholars such as Agnès Celle (2009) have studied how the social dimensions of hearsay are rendered linguistically, drawing attention to hearsay adverbs such as *allegedly*, *reportedly*, and *supposedly* as particularly salient markers. According to Celle, hearsay adverbs are both modal and recipient-oriented in function, and operate uniquely by detaching the speaker from his or her utterance. Unlike other epistemic adverbs, they do not serve to express the speaker's judgment, but rather to neutralize the assertion, preventing the speaker from being held accountable to a judgment of his or her own. This has the important effect of insulating the assertion at hand in a layer of suspended belief.

The hearsay effect reported in the previous chapter documents a correspondence between the use of hearsay adverbs in a tweet and the level of engagement marshaled by that tweet. On #ufotwitter, we observe that tweets with *allegedly* in the text receive, on average, 10.214 times more retweets and 8.529 times more replies than the corpus-wide average. Tweets with *supposedly* receive 4.692 times more retweets and 4.038 times more replies. Tweets with *reportedly* do not exhibit the hearsay effect – they receive a slightly less-than-average number of replies and retweets (as seen in 3-12) – although this may have to do with Celle's observation that, of the three hearsay adverbs, *reportedly* is the most neutral. On platforms like Twitter, where value lies not just in content but in its distribution as well, such a conspicuous increase in engagement suggests that explicitly "weak" formulations of evidence develop a certain kind of power as they are shared and engaged with. This is of particular consequence especially in light of current debates on misinformation and how it proliferates on social media. In the following sections, we consider how extensible the hearsay effect is in two ways. First, we attempt to characterize the relationship between the hearsay adverb and the engagement level using a computational approach. Second, we examine the generalizability of the hearsay effect by adapting it to different discursive contexts on Twitter. We construct multiple corpora of tweets centered around various topics, from the 2021 coup in Myanmar, NBA trade and playoff predictions, and the 2022 Russian invasion of Ukraine, and compare the hearsay effect as measured in each corpus.

4.2 Causality

The hearsay effect as observed in the #ufotwitter corpus is distinctive, but its explanatory power is unclear. Put simply, the hearsay effect documents a correspondence between hearsay adverbs and engagement, but there is no way of telling if the engagement is a direct effect of the mobilization of the hearsay adverb. As explicated by the fundamental problem of causal inference, demonstrating a causal link between the use of a hearsay adverb and the level of engagement received would require that we observe the level of engagement drawn from each tweet in the corpus – posted by the exact same author, at the exact same time, seen by the exact same audience as before – just without the use of a hearsay adverb in the text. This is, of course, infeasible.

Here we describe a research design that approaches this problem from a different angle: using natural language processing (NLP) techniques, can we develop an algorithm that can (somewhat) robustly predict the level of engagement for a tweet, given the text? Then, can we discern noticeable differences in the predicted levels of engagement for tweets with and without the hearsay adverb present in the text?

In this section we introduce the model, the training schema, and results from the first few iterations of algorithmic development. Due to time and compute constraints, this component of the project was ultimately deprioritized before completion, but we offer an analysis and discussion of the existing results for future consideration.

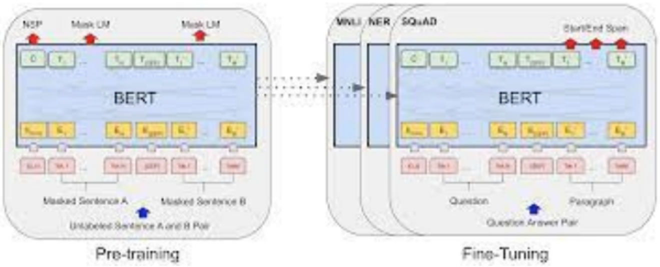


Figure 4-1: Pre-training and fine-tuning with BERT

4.2.1 BERT

To computationally probe the function of hearsay adverbs in the context of a tweet, we “train” a model to recognize patterns between text and a real number “target” value Y (an engagement metric, such as the number of replies) until it becomes reliable enough to predict engagement values for text it has not seen before. We use Bidirectional Encoder Representations from Transformers (BERT) to develop our algorithm (Devlin et al. 2019).

BERT belongs to a family of natural language processing systems called large language models (LMs), which have received considerable media attention for state-of-the-art performance on a number of classic NLP benchmarking tasks. These models tend to have extremely large numbers of parameters – BERT has 340 million – and are used as the backbone of many “transfer learning” schemes in NLP.

Transfer learning is a powerful technique in which a model is first pre-trained on a data-rich, general task before being fine-tuned on a more specific downstream task. During the pre-training phase, the goal is to show the model many examples of natural language while iteratively adjusting the values of its parameters until a baseline level of performance is achieved. BERT is pre-trained on two tasks: masked language modeling and next sentence prediction. In masked language modeling, a small subset of the tokens (15% in the original BERT paper) in the input text are hidden from the model, and the likelihood of a token given its surrounding context is predicted. In next sentence prediction, the model is tasked with predicting how likely a candidate sentence is to follow a given input sentence. Because both of these tasks are unsupervised (meaning that they do not require labels), BERT and most other

LMs are pre-trained on huge quantities of input data, usually taken from the internet (for example, one of BERT's sources for pre-training data is the entirety of English Wikipedia). BERT's success is often attributed to these features of the pre-training process, which allow the model to "learn" numerical vector representations of words that are relatively rich and contextual.

In the fine-tuning step, the model is trained on smaller, more specifically-formatted datasets to do the desired form of prediction. Data for training and validating models for common downstream tasks such as text classification, semantic similarity, and question answering can be found in industry-standard benchmarks such as General Language Understanding Evaluation (GLUE) or Stanford Question Answering Datasets (SQuAD 1.1 and 2.0), but fine-tuning can be adapted for any custom NLP task with a sufficiently high-quality dataset.

The general idea is to show enough examples to the LM in the training process, and program it to consider the surrounding context when making predictions to solve the pre-training tasks, so that the model's many parameters adjust to a mapping of words (or phrases) to a high-dimensional vector space that captures a significant amount of complexity in the language. This is, of course, simply another formulation of pattern matching – critical scholars have rightfully cautioned against mistaking LM-driven performance gains for actual natural language understanding, and argued that the models are better off understood as "stochastic parrots," susceptible to harmful biases encoded in the training data and prone to basing predictions on spurious linguistic cues (Bender et al. 2021; Bras et al. 2020; Niven and Kao 2019). This thesis acknowledges and attempts to leverage the brittleness intrinsic to LMs to inform the model design. That is, instead of training a model to feign understanding of the linguistic function of hearsay adverbs, we use the model to learn a mapping from text to a target value (in this case, the engagement metric) following the statistical distribution of the #ufotwitter corpus. Then, we test the extent to which the hearsay adverb acts as a "cue" for the model in making its predictions. Although LMs are notoriously uninterpretable, we posit that this model design will give us a useful proxy for assessing the "role" of the hearsay adverb in predicting engagement.

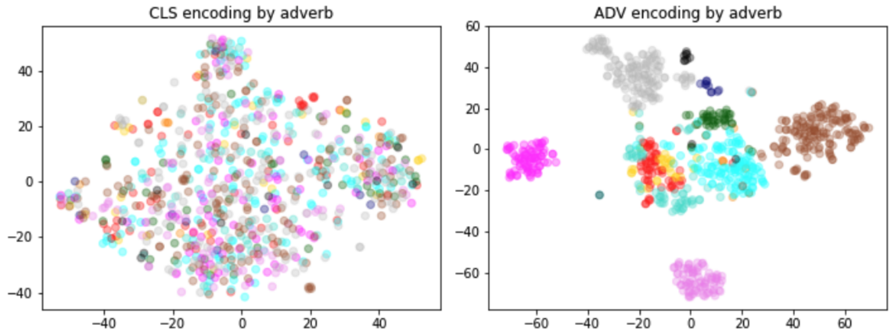


Figure 4-2: Sentence (CLS) embeddings and adverb token (ADV) embeddings for the annotated #ufotwitter dataset, grouped by adverb

4.2.2 Fine-tuning BERT

Here we present four experimental designs for fine-tuning BERT on our custom #ufotwitter dataset. The experiments can be grouped under two main umbrellas: fine-tuning with text classification and fine-tuning with semantic similarity. All experiments are conducted using the Huggingface Transformers library (Wolf et al. 2020) on an NVIDIA T4 GPU.

Before we fine-tune, however, we first assess the baseline performance of BERT. Because the LM has been pre-trained on a large quantity of English language text, but not on tweets specific to #ufotwitter, we expect it to be fairly good at grouping semantically-similar tokens together, but not so good at grouping semantically-similar tweets. To test this, we extract sentence embeddings (represented by the CLS token in the BERT model) and individual token embeddings for the epistemic adverbs present in our dataset using pre-trained BERT model weights. The embeddings are 768-dimensional; for visualization, the 2D t-SNE embeddings are plotted in 4-2 (Maaten and Hinton 2008). As expected, baseline BERT does not recognize any coherent groupings of tweets, but each adverb seems to occupy discrete embedding spaces, with some mixture.

4-3 shows the k-means clustering of the adverb token embeddings when k=6. The composition of the mixtures within each cluster, as documented in 4.1, suggests that baseline BERT maps epistemic adverbs into high-dimensional vector space in a way that is roughly consistent with the sociolinguistic scholarship reviewed in this thesis.

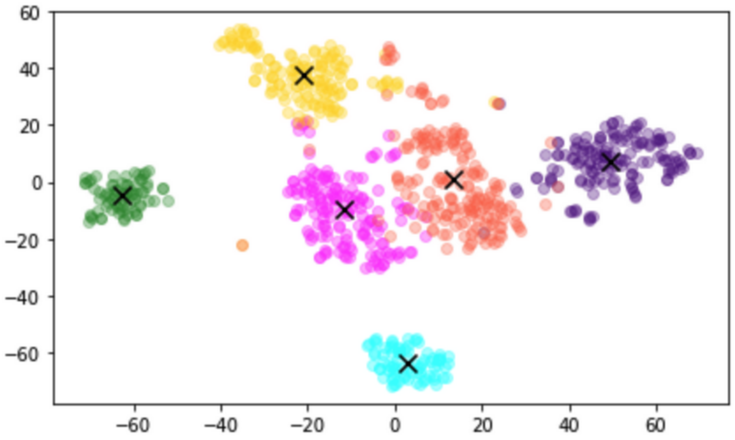


Figure 4-3: k-means clustering of adverb token (ADV) embeddings, $k = 6$

Note that adverbs *clearly* (indigo), *basically* (cyan), and *certainly* (green) have no mixture in their clusters. However, the magenta cluster is composed of hedges and hearsay verbs, including *allegedly*, *reportedly*, and *supposedly*. With this baseline level of performance in mind, we proceed with the fine-tuning design.

Adverb	Cluster	Number of tokens
Clearly	Indigo	148
Basically	Cyan	79
Supposedly	Magenta	58
Allegedly	Magenta	58
Reportedly	Magenta	27
Possibly	Gold	140
Certainly	Green	75
Apparently	Tomato	112
Seemingly	Tomato	44

Table 4.1: Variation in descriptive categorizations of epistemic adverbs

The first set of fine-tuning tasks focus on *text classification*. We experiment with three flavors of text classification: binary classification, multiclass classification, and ordinal regression. For all experiments, we fine-tune the model on a small amount of data, compute the embedding for each tweet in the annotated #ufotwitter dataset, and repeat the dimensionality reduction process described above to test if the model recognizes coherent groupings of tweets based on the epistemic adverb they mobilize.

To fine-tune with binary classification, we create a small training dataset with the

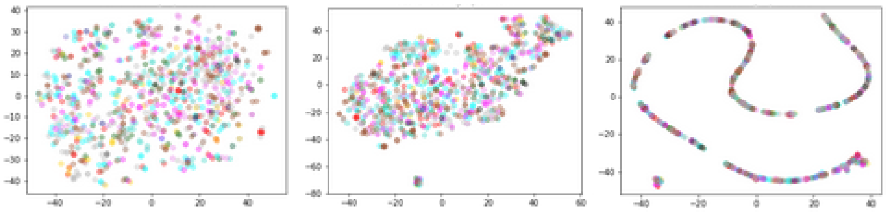


Figure 4-4: t-SNE tweet embeddings after fine-tuning for text classification

binary class labels {*evidential*, *non-evidential*}. We take all 800 samples from the annotated #ufotwitter dataset as positive (evidential) samples, and randomly sample and annotate tweets from the rest of the #ufotwitter corpus until we accumulate 500 negative samples. It is worth noting that the annotation step is difficult to scale, especially without clear markers of "non-evidentiality". With this small training set, BERT is fine-tuned using the default learning rate and optimizer for 10 epochs[1].

After fine-tuning, the CLS token is extracted for each tweet and projected back into 2-dimensional space. The plots are shown in 4-4: from left to right, embeddings are extracted from the fine-tuned binary classification model, multiclass classification model, and ordinal regression model. They suggest that text classification is not a viable fine-tuning task for the problem at hand. For simple classification tasks (like binary classification), it is impractical to develop a labeled training set that is large and robust; for classification tasks that use existing attributes in the corpus (like the number of replies) for supervision, the learning task becomes intractably complex and computationally expensive.

Motivated by previous work on learning state-of-the-art sentence embeddings with LMs like BERT, we try another method of fine-tuning using semantic similarity. According to Reimers and Gurevych (2019), BERT CLS embeddings are mapped in a vector space that is unsuitable for usage with common distance-based similarity measures (which is what we are looking for in our clustering-based tests). They propose a new architecture, SentenceTransformers, that is fine-tuned on semantic similarity tasks. SentenceTransformers compute dense vector representations at the

[1] The default learning rate is 5e-05 and the default optimizer is AdamW. An epoch is one complete pass through the training data.

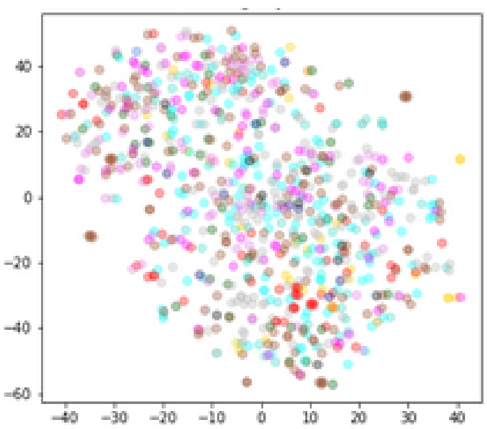


Figure 4-5: Tweet embeddings extracted from SentenceTransformers with DistilRoBERTa backbone (no further fine-tuning)

sentence-level such that similar text is pushed together "closer" in the vector space.

As shown in 4-5, some use cases (such as this one) require further fine-tuning. However, fine-tuning a SentenceTransformer model when the training set consists only of single (text, label) pairs, as in the #ufotwitter corpus, is a non-trivial task. This is because the SentenceTransformer architecture is itself a modification of a transformer that uses Siamese and triplet networks for its own fine-tuning step. One of its innovations is in taking advantage of pre-existing NLP training data with supervised sentence pairs or triplets, and selecting objective functions based on the type of annotation that is available – effectively showing the model examples of groups of text that are "similar" or "dissimilar". For example, when fine-tuning on the Stanford Natural Language Inference (SNLI) corpus (Bowman et al. 2015), which is comprised of sentence pairs labeled for entailment, contradiction, and semantic independence, SentenceTransformers uses the classification objective

$$\text{softmax}(W_t(u, v, |uv|))$$

where u, v are sentence embeddings and the trainable weight W_t is defined $W_t \in R^{3n \times k}$. When fine-tuning on STS-B, the Semantic Textual Similarity benchmark, which provides sentence pairs labeled with a human-annotated real-value similarity score, SentenceTransformers uses a regression objective defined by the mean-square error loss of the cosine similarity between the embeddings of the sentence pairs.

When fine-tuning on weakly-labeled triplets such as those found in the Dor et al. (2018) Wikipedia sections dataset, which hinges on the assumption that sentences found in the same section of Wikipedia are more likely to be semantically similar than sentences found in different sections, SentenceTransformers uses a triplet loss objective function

$$\max(||s_a s_p||||s_a s_n||+, 0)$$

where each triplet consists of an "anchor" sentence a, a positive (similar) sentence p, and negative (dissimilar) example n. The loss is computed using the embeddings s_a, s_p, and s_n.

However, with the #ufotwitter corpus, we have no immediate way of getting similar annotations. In order to fine-tune SentenceTransformers with our custom data, we must mine our own triplets.

Mining triplets is a complex problem and has primarily been implemented for computer vision studies training deep convolutional neural networks (not the transformer architectures that have emerged for solving NLP tasks); thus, the work outlined herein is highly experimental. As Hermans et al. (2017) observe, the number of possible triplets grows cubically as the dataset gets larger, making it necessary to only mine "good" triplets – triplets that are sufficiently hard (otherwise training will quickly stagnate) but not too hard (otherwise training will be unstable). They recommend the following "Batch Hard" method to compute the loss:

$$\mathcal{L}_{BH}(\theta;X)=\sum_{i=1}^{P}\sum_{a=1}^{K}[m+\max_{p=1\ldots K} D(f_\theta(x_a^i), f_\theta(x_p^i)) - \min_{\substack{j=1\ldots P,\\ n=1\ldots K,\\ j\neq i}} D(f_\theta(x_a^i), f_\theta(x_n^j))]$$

such that for a mini-batch X and each sample a within the batch, the hardest positive and hardest negative samples within the batch are mined to form the triplets. The Batch Hard triplet loss described in Hermans et al. (2017) is primarily used for facial re-identification applications, where the notion of the "same face" or "different

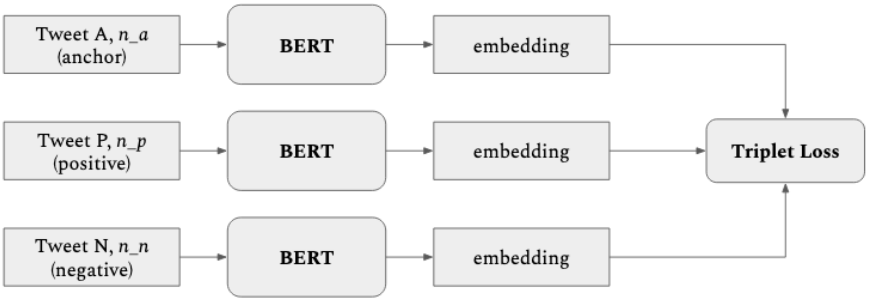


Figure 4-6: Triplet network architecture

face" is more likely to be defined within the dataset (although this is in and of itself problematic, cf. Keyes 2018; Stevens and Keyes 2021). To adapt it to the #ufotwitter corpus, which has no prerecorded indicator of similarity, we construct an artificial class label based on the level of engagement associated with each tweet. We draw "similar" pairs a and p from the set of tweets i that fulfill the condition $y_i - \mu \leq \sigma$, where y_i is the target label (number of replies), μ is the corpus-wide mean, and σ is the corpus-wide standard deviation. Negative samples n are drawn from the set of tweets that do not fulfill this condition. By fine-tuning a standard triplet network structure (see 4-6 for a schematic) on these triplets, the algorithm in effect "pushes together" tweets that have low levels of engagement in the high-dimensional vector space, and "pushes apart" the tweets that have high levels of engagement.

4.2.3 Results and limitations

Training with mined triplets is tricky to stabilize. Here we present the result of several different experimental setups varying the learning rate and batch size, as well as the normalization of the target variable y_i.

Training without normalization of the target variable tends to be unstable, as depicted in 4-7. We turn to two common methods of normalization: logarithmic transformation, with

$$y'' = log(y)$$

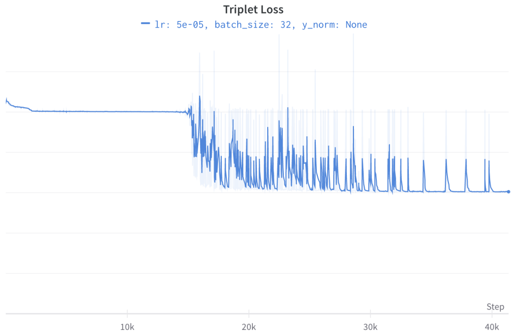


Figure 4-7: Fine-tuning triplet loss with default learning rate and batch size, and no normalization of the target variable

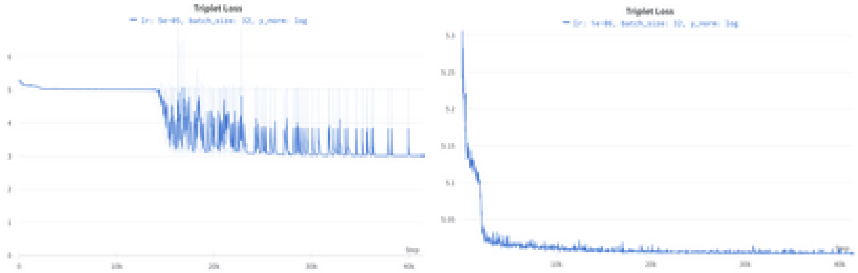


Figure 4-8: Fine-tuning triplet loss with log normalization of the target variable

and min-max scaling, with

$$y' = \frac{y - y_{min}}{y_{max} - y_{min}}$$

As shown in 4-8 and 4-9, the learning rate must be adjusted in order for training to stabilize with either normalization scheme. However, even after adjusting the learning rate, the loss converges at a relatively high value in both cases (suggesting that the model has not "finished learning" from the training data in the fine-tuning step).

We also experiment with a larger batch size of 64, but we run into memory issues with the compute target before convergence, as shown in 4-10.

We do not try other training configurations that would require more computational power or training time. Instead, we turn to the problem of assessing the usefulness of these fine-tuned models for our original task: comparing tweet embeddings with and without key epistemic adverbs. 4-11 shows that the embeddings extracted from an incompletely fine-tuned model are indeed an improvement from the CLS tokens extracted directly from fine-tuned BERT – they can be differentiated by the number of

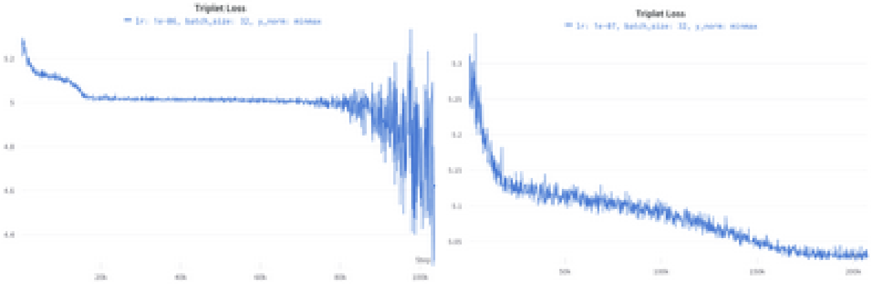


Figure 4-9: Fine-tuning triplet loss with minmax normalization of the target variable

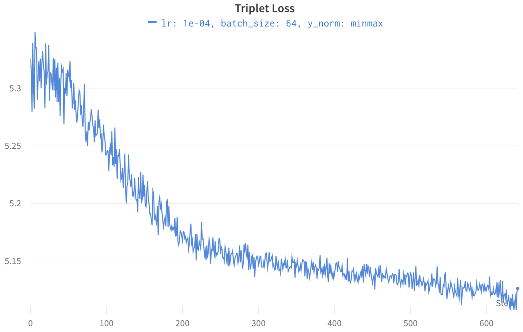


Figure 4-10: Fine-tuning triplet loss with larger batch size

replies associated with the tweet. However, in 4-12 we see that the embeddings have overfit to the target value, and are not consistently sensitive to changes in the text. We plot the number of replies y_i for a given tweet i against the Euclidean distance d_i between the tweet embedding s_i and modified tweet (without the adverb) embedding t_i, and find a significantly wider spread for d_i when engagement is low, and almost zero variation in d_i when engagement is high. This could be attributed to imbalances in the training dataset, or unstable or incomplete training in the fine-tuning step. Ultimately, however, we accept that the marginal return to further algorithmic development in terms of developing our understanding of evidentiality is in fact diminishing, and shift our focus to other methods of apprehending the extensibility of the hearsay effect. In the following section we attempt to measure the hearsay effect in other discursive communities on Twitter, and discuss the generalizability of the hearsay effect across different epistemes and sites of controversy online.

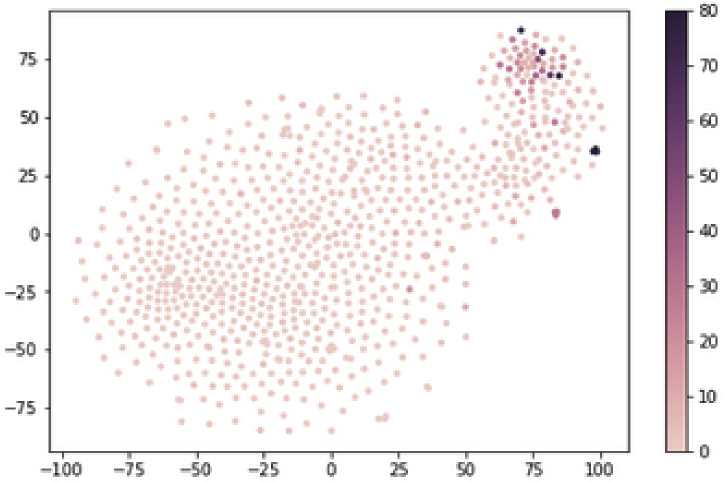


Figure 4-11: 2D t-SNE embeddings from fine-tuned SentenceTransformers model fine-tuned grouped by number of replies)

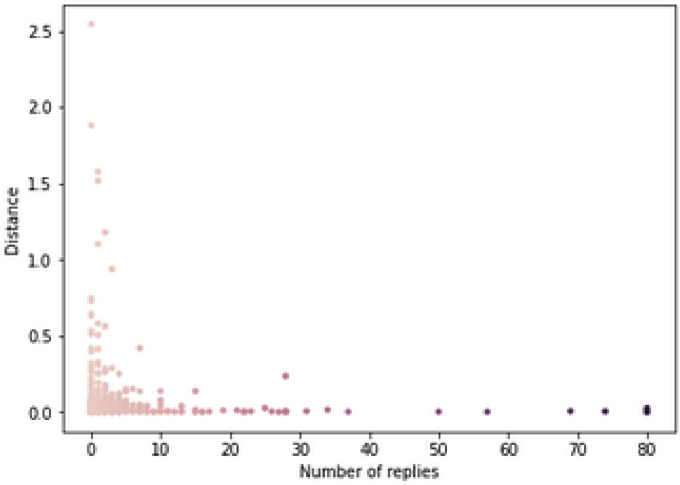


Figure 4-12: Variation in Euclidean distance between embeddings for tweets with and without key epistemic adverbs in the #ufotwitter corpus

4.3 Generalizability

To assess its generalizability, we measure the hearsay effect in other discursive communities on Twitter that engage in highly evidential talk.

4.3.1 Formula

The hearsay effect H is measured as follows. For a given corpus X, we define r_i to be the number of replies for tweet i, and t_i to be the text of tweet i. We also define A as the set of tweet indices where the hearsay adverb "allegedly" is found in t_i. Then,

$$R_A = \frac{\sum_{i \in A} r_i}{\mathbf{card}(A)}$$

$$R_X = \frac{\sum_{i \in X} r_i}{\mathbf{card}(X)}$$

$$H = \frac{R_A}{R_X}$$

H is easy to compute, but it is only useful if the boundaries of the discursive community represented by X are well-defined. For the #ufotwitter corpus, this was established through a process of manual annotation after ethnographic study. We do not want to have to manually annotate every test case used for our generalizability assessment, so we leverage several features of the Twitter API to inform the construction of our test corpora. This method for data collection is summarized below.

4.3.2 Test data

There are two main criteria for developing a good test case for the hearsay effect:

(1) The discourse involves discussing and debating evidence, and

(2) The discourse is centered around a coherent topic

As discussed in previous chapters, there is no single, infallible marker of evidential discourse that can be used to perfectly satisfy (1). Instead, we use the fact that the hearsay adverb *allegedly* is highly indexical of (1) to bound a search for salient hashtags and keywords that satisfy (2). Concretely, this translates into a four-step process, enumerated here:

(i) Mine tweets that mobilize the hearsay adverb *allegedly* in the text;

(ii) Cluster the tweets by topic using topic modeling techniques;

(iii) Extract salient hashtags and keywords from each cluster to develop topical search queries; and

(iv) Collect tweets from each topical search query to construct the test corpora

Of course, this is a highly imperfect heuristic – the topics generated by this method are constrained by the tweets collected in the first phase with the *allegedly* keyword search, which is not reliably representative of all evidential talk on Twitter. Moreover, tweets in the test corpora that do *not* use the hearsay adverb may not even be evidential in nature. Nevertheless, we implement the data collection process in an effort to gather a wide array of candidate test case topics for consideration.

We collect over 3 million unique tweets from an eighteen-month period ranging from August 2020 to February 2022, using *allegedly* as the search query. We implement BERTopic to cluster the tweets. BERTopic proceeds in four steps: first, the tweets are mapped to a high-dimensional vector space using BERT; the dimensionality is then reduced with uniform manifold approximation and projection (UMAP) to make clustering computationally tractable (McInnes, Healy, and Melville 2020). Dense clusters are identified using the HDBSCAN algorithm (Campello, Moulavi, and Sander 2013). Within each cluster, topics are surfaced using c-TF-IDF, a class-based variant of term frequency-inverse document frequency (Grootendorst 2020).

The most salient topics, however, are not necessarily good candidates for translating into test cases. This can be seen more concretely in 4.2 and 4.3, which record the clusters as organized by size and reply rate. Building a search query that can satisfy criteria (1) and (2) is difficult with some of these keywords, which are often too general or multivalent in their use on Twitter. To work around this, we also surface the most salient hashtags, as documented in 4.4, and annotated entities, shown in 4.5.

Topic	Tokens
1	Vaccine, vaccinated, vaccines, unvaccinated
2	Masks, mask wearing, mask wear, mask
3	Album, song, songs, music
4	Biden, Biden won, votes, Joe Biden
5	Palestinian, Palestinians, Hamas, Israeli
6	Package, Delivered, Parcel, Delivery
7	Tories, Labour, Tory, MPS
8	Lebron, Lakers, Harden, Kyrie
9	Laptop, Hunter, Repair, Hard drive
10	Britney, Britneys, Jamie, Spears

Table 4.2: Ten largest BERTopic clusters in the 3M *allegedly* sample

Topic	Tokens
22	Trisha, Ethan, Trish, Trishas
16	Epstein, Epsteins, Jeffrey Epstein, Jeffrey
13	Racist, Racism, White people, Racists
27	Gaetz, Matt Gaetz, Matt, Greenberg
52	Black man, Black people, White people, White
61	FDA, Pharma, Doses, Pharmacy
34	Dog, Dogs, Puppy, Pup
8	Lebron, Lakers, Harden, Kyrie
45	Scam, HMRC, Calls, Number
7	Tories, Labour, Tory, MPS

Table 4.3: Clusters with the highest reply rates in the 3M *allegedly* sample

Hashtags
whatshappeninginmyanmar
smartnews
news
crimesagainsthumanity
endsars

Table 4.4: Top five most salient hashtags in the 3M *allegedly* sample

Using this information, we manually select[2] eight topics of interest and craft corresponding search queries for each topic, as shown in 4.6. All queries are run for the

[2] A highly subjective procedure in which we subsampled and inspected a small amount of tweets

Hashtags
Trump Biden US Florida FBI

Table 4.5: Top five most salient annotated entities in the 3M *allegedly* sample

same eighteen-month period as above, with the exception of the last query, which tracks emergent Twitter conversation on the Russian invasion of Ukraine from February 24, 2022 to March 14, 2022. The hearsay effects measured within each of the test corpora are documented and discussed in the following section.

Topic	Search Query
Myanmar coup	(#WhatsHappeningInMyanmar OR #WarCrimesOfJunta) lang:en -is:retweet
Covid-19	(coronavirus OR covid19 OR covid-19 OR covid OR pandemic) (plot OR chart OR map OR dashboard OR vis OR viz OR visualization) lang:en -is:retweet
Britney Spears	(#freebritney OR (britney spears)) lang:en -is:retweet
Jeffrey Epstein	epstein lang:en -is:retweet
Hunter Biden	(hunter biden) OR hunterbiden) lang:en -is:retweet
Matt Gaetz	gaetz lang:en -is:retweet
NBA	(#nbatwitter OR #nba) lang:en -is:retweet
Ukraine	(ukraine OR russia OR putin) lang:en -is:retweet

Table 4.6: Search queries for test corpora

4.3.3 Results

The hearsay effect as computed for the test corpora is relatively muted, especially in comparison to the #ufotwitter corpus, in which $H = 8.529$ – more than double the highest posted value of H in 4.7. There are a few possible explanations for this.

On one hand, it could be that #ufotwitter is unique in the ways in which members circulate and negotiate evidence, and that the hearsay effect is relevant only in this

from each candidate topic before determining whether or not they would make a good candidate. For example, although "Covid" and "vaccine" are highly salient keywords surfaced by the data collection methodology, we find in their respective subsamples that many of the tweets are non-evidential.

singular discursive community.

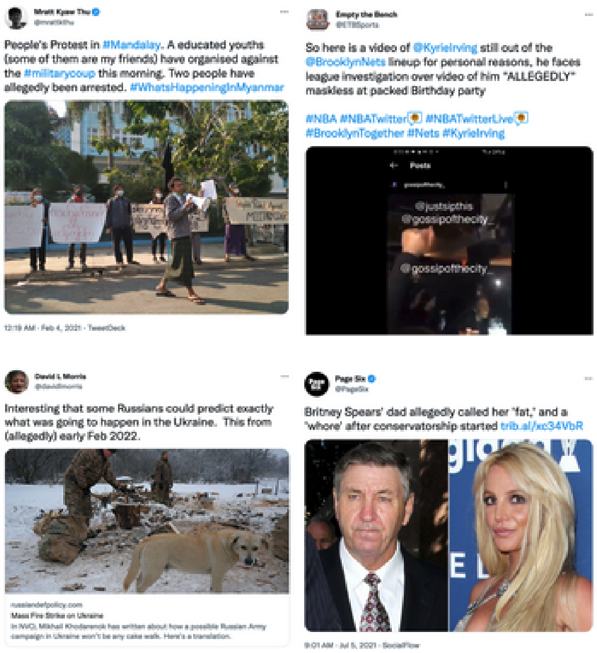
It could also be that the test corpora capture a substantively different kind of discourse than what is found on #ufotwitter. Several illustrative examples are shown in 4-13. Tweets coming from journalists commenting on breaking news in war zones or stories from pop culture, or otherwise mirroring an expert evidential register in which hearsay adverbs are conventionally deployed for professional legitimization, are distinct from instances of a discursive community coming together to adjudicate evidence online (as is the case with #ufotwitter). With this difference in form, it makes sense that the function – the circulatory effect denoted by H – is not directly comparable.

Topic	H
Myanmar coup	1.186
Covid-19	2.157
Britney Spears	1.484
Jeffrey Epstein	1.304
Hunter Biden	2.396
Matt Gaetz	1.693
NBA	1.686
Ukraine	1.290

Table 4.7: Hearsay effects in the test corpora

Can we find this form of evidential discourse elsewhere? Motivated by recent coverage of open source intelligence (OSINT) analysis of the ongoing war in Ukraine (Schwartz 2022; Vick 2022), we recompute the hearsay effect on a subsample of the Ukraine test data. The subsample is defined by extracting tweets from the Ukraine test data that are "related" to a coarsely-defined set of users central to the OSINT Twitter community (Appendix B). A tweet is included in the subsample if it is authored by, replies to, or mentions an account in the OSINT user set. In the newly-constructed OSINT corpus, we observe a significant magnification in H from 1.290 to 3.117.

In hindsight, the parallels between OSINT and ufology, and the prominence of the hearsay effect in their respective online communities, is unsurprising. OSINT refers to amateur intelligence research conducted using free and publicly accessible



Mratt Kyaw Thu
@mrattkthu

People's Protest in #Mandalay. A educated youths (some of them are my friends) have organised against the #militarycoup this morning. Two people have allegedly been arrested. #WhatsHappeningInMyanmar

12:19 AM · Feb 4, 2021 · TweetDeck

Empty the Bench
@ETBSports

So here is a video of @KyrieIrving still out of the @BrooklynNets lineup for personal reasons, he faces league investigation over video of him "ALLEGEDLY" maskless at packed Birthday party

#NBA #NBATwitter #NBATwitterLive
#BrooklynTogether #Nets #KyrieIrving

David L Morris
@davidlmorris

Interesting that some Russians could predict exactly what was going to happen in the Ukraine. This from (allegedly) early Feb 2022.

russiandefpolicy.com
Mass Fire Strike on Ukraine
In NVO, Mikhail Khodarenok has written about how a possible Russian Army campaign in Ukraine won't be any cake walk. Here's a translation.

Page Six
@PageSix

Britney Spears' dad allegedly called her 'fat,' and a 'whore' after conservatorship started trib.al/xc34VbR

9:01 AM · Jul 5, 2021 · SocialFlow

Figure 4-13: Sample tweets from the test corpora

information to track things like military activity, flight patterns, arms flows, and other strategic security issues. Indeed, many ufologists consider their own work to be a form of OSINT, especially those who are motivated by a desire to extract the truth from underneath a veil of suspicious state secrecy. OSINT researchers aim to do just that. Using open-access data from security video feeds, free satellite imagery services, and social media, they publicly conduct the type of work that intelligence agencies do behind closed doors, frequently collaborating with each other to trade tips on where to find information and how to analyze it. The community, which got its start in 2011 operating on the fringes of the Syrian civil war, has grown massively over the years – one of the more popular Discord servers where researchers gather, Project Owl, has more than 30,000 members (Schwartz 2022). Now, OSINT enjoys more mainstream recognition. Perhaps most famously, the investigative outlet Bellingcat (a heavy-handed allusion to resourceful mice tying a bell to a cat) received significant media attention in 2014 after proving that a Russian surface-to-air missile was responsible for the downing of MH17, a Malaysian passenger jet flying over eastern Ukraine. More recently, OSINT has exploded into the public consciousness after circulating evidence documenting the scale of Russian military activity on Ukraine's borders months before the invasion, and offering quick entrypoints into debunking Russian disinformation and propaganda about the war (Vick 2022; Schwartz 2022).

Bellingcat founder Elliot Higgins, who maintains that the "response to disinformation is transparency" (Vick 2022), summarizes the OSINT ethos well. But the conspicuous hearsay effect detected in OSINT Twitter discourse draws attention to a part of the story that often goes unnoticed. Transparency, for all its ideological trappings, is not actually an end in itself. When it is weak evidence that is circulated and otherwise engaged with, the effect is more similar to the diffusion of public secrets or the revelation of concealment – where secrets are socially mobilized as a form of sociocultural capital without dispersing restricted knowledge (Jones 2014). The legitimizing effect of evidential buck-passing in these contexts underscores the importance of understanding how qualities of entextualization modulate the perlocutionary effects of evidential claims online, and raises new questions on the strength of weak evidence.

4.4 The strength of weak evidence

So far in this chapter, we have examined the hearsay effect through multiple lenses. In the first section, we develop an understanding of the linguistic production of hearsay evidence, which is uniquely recipient-oriented. This allows the speaker to animate data-as-evidence while refusing to commit to a point of view. The high engagement levels documented by the hearsay effect on #ufotwitter suggest that it is indexical of a certain genre of evidentiality, where the suspension of belief itself becomes a currency of legitimacy.

Although it is difficult to pin down if the hearsay effect is directly attributable to the deployment of specific linguistic features, such as hearsay adverbs, we find that we do not need to establish a causal link to begin to explain it. We observe that the hearsay effect is variably salient across different discursive communities on Twitter, the significance of which is twofold. First, it shows that the hearsay effect is not something that is unique to #ufotwitter. Second, it indicates that the hearsay effect is amplified in contexts where institutional trust is low and central epistemic authorities are absent. This is not as trivial of an observation as it might at first seem. The correspondence developed by the hearsay effect between linguistic markers (that invite the listener, or recipient, to construct a point of view) and engagement points to important dynamics underpinning the strength of weak evidence within contested sites of knowledge production. If anything, the hearsay effect demonstrates that there is a lot to learn from focusing on the embeddedness of information in interactional and communicative practices. This, in turn, bears profound consequences for the study of misinformation and disinformation more broadly. Indeed, making sense of the strength of weak evidence may be a critical step forward in understanding and stabilizing the volatile information ecologies that are characteristic of social media platforms today.

Chapter 5

Conclusion

This thesis is a story about an adverb, *allegedly.* But it is also a story about data, evidence, and knowledge – and how people talk about information as it circulates online. Starting with a deceptively simple question – can the data speak for itself? – we first unravel how data is animated and taken up as evidence through a fundamentally interactional and dialogic process. In our case study of #ufotwitter, we see how nuanced and diffuse evidential production can be, especially in the absence of epistemic or institutional authority. We also see, empirically, how inflections in the way that data is entextualized shape the social relations that modulate if and how that data is taken up as evidence. We observe that formulations of data-as-evidence that are indexical of hearsay are correlated with high levels of engagement and circulation, and introduce the hearsay effect to characterize this phenomenon. By testing the hearsay effect on different discursive communities, we find that the hearsay effect is variably salient and indexical of important dynamics underpinning the strength of weak evidence within contested sites of knowledge production.

There are multiple possible interpretations of the strength of weak evidence on social media. The first is the simplest, and perhaps the most complementary to the ways in which misinformation is most commonly framed in both the media and the academy. In this view, the strength of weak evidence is ascribed to certain epistemological features of the discourse community that perpetuate the circulation of low-quality information online. These features may be related to a lack of reflexivity or media literacy from the audience, or an incentive system that rewards the production of misinformation with political or economic gain. This in fact seems to be the dominant characterization of what constitutes the “infodemic” in the modern day. However, the hearsay effect substantiates other narratives about misinformation as well. When the practice of evidentiality is heavily politicized, as it is in the school reopening tweet discussed in Chapter 1, weak evidence may function as an invitation for heightened scrutiny, spurring even *more* critical engagement. Moreover, in the absence of institutional trust and authority, the hearsay adverbs that construct "weak evidence" may actually lend the data credence that boosts uptake and engagement. Yet another interpretation suggests that the simultaneous accumulation of legitimacy and refraction of accountability motivate the reanimation of circulation of evidence, as it is what sustains the “longest path” to sites of privileged, restricted knowledge. And in contexts such as in #ufotwitter, the imaginative power of the hearsay effect, which allow multiple, mutually exclusive possibilities to coexist (Celle 2009), could be what drives engagement. This in turn allows for the continual construction and proliferation of possible worlds where multiple interpretations of the evidence could exist simultaneously (Hanks 2016).

One or more of these interpretations could be operative for any given discursive community. It follows that making sense of the hearsay effect, then, requires a highly nuanced understanding of the sociocultural context, as the metric itself does not carry much explanatory power. While this paints a rather bleak picture of our ability to automatically identify and correct the manifold conditions that animate misinformation (or, for that matter, information) online, it introduces new points of departure for thinking about the role of language in mediating perceptions of truth and reality in multimodal, multidimensional landscapes of evidentiality. In an increasingly fragmented, yet cacophonous online public sphere, it is necessary to develop new methods and frameworks to reconcile big data with “thick description,” so that the materialization of phenomenologically-rich concepts of stance and evidentiality can be analyzed at scale. This work is a small step in that direction.

Appendix A

LDA Visualization

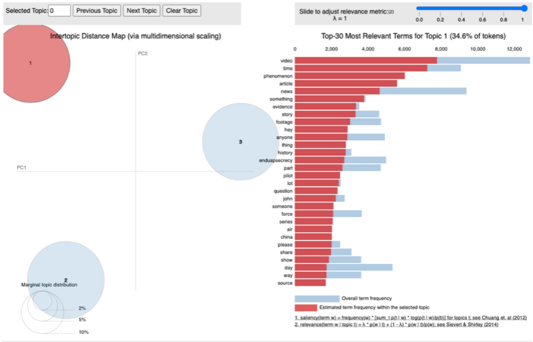

Selected Topic: 0
Previous Topic
Next Topic
Clear Topic
Slide to adjust relevance metric:(2)
λ = 1
0.0
0.2
0.4
0.6
0.8
1
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
1
2
3
Marginal topic distribution
2%
5%
10%
Top-30 Most Relevant Terms for Topic 1 (34.6% of tokens)
0
2,000
4,000
6,000
8,000
10,000
12,000
video
time
phenomenon
article
news
something
evidence
story
footage
hey
anyone
thing
history
enduapsecrecy
part
pilot
lot
question
john
someone
force
series
air
china
please
share
show
day
way
source
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure A-1: Top 30 most salient terms in Topic 1 of the LDA topic model

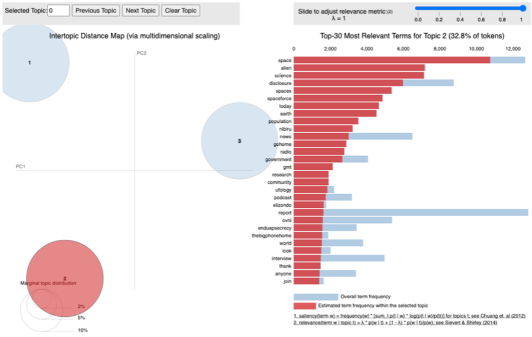

Selected Topic: 0
Previous Topic
Next Topic
Clear Topic
Slide to adjust relevance metric:(2)
λ = 1
0.0
0.2
0.4
0.6
0.8
1
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
1
2
3
Marginal topic distribution
2%
5%
10%
Top-30 Most Relevant Terms for Topic 2 (32.8% of tokens)
0
2,000
4,000
6,000
8,000
10,000
12,000
space
alien
science
disclosure
spacex
spaceforce
today
earth
population
nibiru
news
gohome
radio
government
gmt
research
community
ufology
podcast
elizondo
report
ovni
enduapsecrecy
thebigphonehome
world
look
interview
thank
anyone
join
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure A-2: Top 30 most salient terms in Topic 2 of the LDA topic model

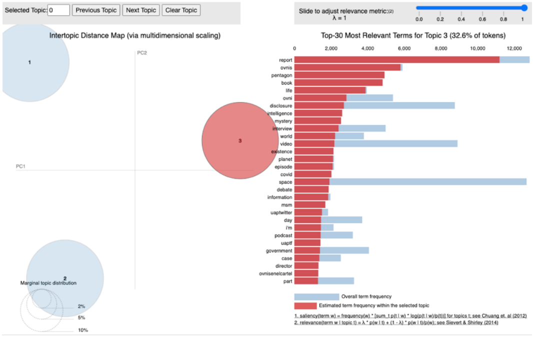

Selected Topic: 0
Previous Topic
Next Topic
Clear Topic
Slide to adjust relevance metric:(2)
λ = 1
0.0
0.2
0.4
0.6
0.8
1
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
1
2
3
Marginal topic distribution
2%
5%
10%
Top-30 Most Relevant Terms for Topic 3 (32.6% of tokens)
0
2,000
4,000
6,000
8,000
10,000
12,000
report
ovnis
pentagon
book
life
ovni
disclosure
intelligence
mystery
interview
world
video
existence
planet
episode
covid
space
debate
information
msm
uaptwitter
day
i'm
podcast
uaptf
government
case
director
ovnisenelcartel
part
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure A-3: Top 30 most salient terms in Topic 3 of the LDA topic model

Appendix B

OSINT Filter

Table B.1: Twitter accounts used to filter the OSINT subsample

Username	User bio
IntelCrab	Open-source intelligence analyst, map nerd, and crabby newsdesk manager for @CBS_42. Political science student at UAB. Retweets are not automatic endorsements.
CovertShores	Defense Analysis, Submarines, #OSINT, illustrations and history. Author of Covert Shores books. Write for @USNINews, @navalnewscom and more. Mostly typos.
AuroraIntel	Team of 3. Bringing you world events as they happen, focusing on the Middle East Region | RTs/Links Endorse | DMs Open
Aviation_Intel	Creator/Editor-In-Chief of The War Zone @ The Drive. Gizmodo Media Group and Time Inc. alum.
PutinIsAVirus	Doesn't like Russian supreme leader very much, extremely interested in OSINT and tries to be active about it + Drones & Pasta
Gerjon_	Aircraft Tracker | Image Analyst | lyushin Il-76 fan | Quoted by U.N. PoE, Al Jazeera, BBC Africa Eye,

Continued on next page

Table B.1 – continued

Username	User bio
	NY Times et al. | DM open - feel free!
obretix	I like brave people who are able to think for themselves. geospatial IT – OSINT/IMINT
thewarzonewire	A strong offense for the world of defense.
YorukIsik	BOSPHORUS OBSERVER: Obsessive ship-spotting by the Bosphorus ..- -. .. - . -.. / ..-. — .-. / ..- -.- .-. .- .. -. .
N_Waters89	Digital investigations, Syria, Yemen, now Ukraine. Ex-army. Justice & Accountability @Bellingcat.
TheAviationist	Unmatched independent coverage of military aviation, defense, technology & everything in the middle. Brought to you by David Cenciotti (@cencio4).
steffanwatkins	Canadian Writer, Researcher, Consultant. Big fan of #AIS #ModeS #ADSB #MLAT #ADSBexchange #OSINT. #Disinformation ≠ #Misinformation. steffan.watkins@gmail.com
SyriaCivilDef	We're the Syria Civil Defence (White Helmets), our humanitarian work helps communities prepare for, respond to & recover from attacks. We've saved +120k lives.
COUPSURE	OSINT Investigator at @Cen4infoRes | Satellite imagery nerd | @L_ThinkTank | @lemondefr | Views are my own, RT ≠ endorsement| ko-fi.com/intelcoupsur
search_ish	a how-to website about OSINT research
monty5stars	
OSINT_AVIA	Paracyclist, Aviatikbegeisterter, OSINT Recherearbeit über die Luftfahrt. he/him
FranticGoat	'Punk' Journalist, Executive Editor @thewarzonewire, and defense and security analyst

Continued on next page

Table B.1 – continued

Username	User bio
Przybyszewski_L	Private account. Warsaw, PL | Founder & President - http://abhaseed.org Former research analyst at a military think-tank.
mirek_szczerba	Former social media advisor to the former President of the Republic of Poland LECH WAŁĘSA. Researcher #OSINT #Geolocation #Security #intel #Barca
Strike_dip_com	Geologist loving 3D geological modelling and all that is associated with it, maps, cross sections, interpretations...!
joanne_stocker	Journalist at @Storyful. Video/open source, currently focused on Ukraine. Yogi, runner, Philadelphian. Big fan of dogs. joanne.stocker@storyful.com
Jef_Hilsen	« Les larmes de nos souverains ont le goût salé de la mer qu'ils ont ignorée » Richelieu
Libya_OSINT	Maritime | Security | Oil & Gas | Focus on Studying LNA Armed Groups | DMs Open
AlbinSzakola	Reporter @lorienttoday; interested in the political economy of Lebanon and Syria, logistics and data journalism; former senior analyst @KharonData
il_kanguru	urx67POzqbt$
segioajv1	Architect, worldwatcher, geogeek I like maps, images, Nature & Humanity
IntelKirby	Project owl discord server: https://discord.gg/projectowl
ACLHaynes	Historian so we can have a better future - The cat is named Setzer. To live without hope is to cease to live.
GambLuca	Journalist @IlFoglio_it - The discipline of the written word punishes both stupidity and dishonesty.
Libyancitizen6	ordinary citizen, tweeting about Libyan daily life/geo-politics

Continued on next page

Table B.1 – continued

Username	User bio
	/military + few other things. Anti Turks/ their dogs in Libya: Türklerin mezarlığı.
NATO_MARCOM	Official twitter account for MARCOM - NATO Allied Maritime Command, Public Affairs. Retweet ≠ endorsement.
KomissarWhipla	Malign Russian/Organically Ruthless/Chaos Bomb/ #STRATDELA/#hyperhype/@IMEMO_RAN/ @Russian_Council/Spartak Msk/Views personal/ Odessa will be free
assoulix	Datascientist passionated by defense domain.
n_morse9927	Primarily monitor AIS
selingirit	BBC News http://Instagram.com/selingirit
EliotHiggins	Founder and creative director of Bellingcat and director of Bellingcat Productions BV. Author of We Are Bellingcat.
Osinttechnical	Someone called me a defense journalist once. I try to communicate risk. @TheOsintBunker cohost and editor. Slightly lost American with @UKDefJournal
wondersmith_rae	OSINT Analyst | @TraceLabs Black Badge Winner & MVO | @OsintCurious | @InnocentOrg | http://safeescape.org | @Quiztime | @WileyTech Author | Maritime Obsessed
trbrtc	Visual Investigations at @nytimes. Previously @bellingcat, @airwars. Retired hitchhiker. Learn digital verification in a fun way → @quiztime
ISNJH	Space and Defense reporting, Specializing in GEOINT and 3D reconstruction / Content contributor Jane's Intelligence Review
cencio4	One of the world's most read military aviation bloggers and the man behind @TheAviationist. Journo, IT Sec pro, Computer Eng., Ret. AF officer, speaker, pilot.

Continued on next page

Table B.1 – continued

Username	User bio
bellingcat	Want to support our charity? http://bellingcat.com/donate/ Buy our book "We Are Bellingcat" here: http://bit.ly/2EP09EN Our award-winning podcast series: http://apple.co/36LJURI
Global_Mil_Info	Navigator of the Globe.
StratSentinel	Created by @W5RKB to provide a source for NatSec, IR, and defense breaking news with commentary. Daily newsletter via Patreon: http://patreon.com/StratSentinel

Bibliography

Agrama, Hussein Ali. 2021. "Secularity, Synchronicity, and Uncanny Science: Considerations and Challenges." *Zygon®* 56 (2): 395–415. ISSN: 1467-9744. https://doi.org/10.1111/zygo.12671.

Aikhenvald, Alexandra. 2004. *Evidentiality.* Oxford Linguistics. Oxford University Press. ISBN: 978-0-19-926388-2.

Anand, Pranav, Marilyn Walker, Rob Abbott, Jean E. Fox Tree, Robeson Bowmani, and Michael Minor. 2011. "Cats Rule and Dogs Drool!: Classifying Stance in Online Debate." In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis,* 1–9. Portland, Oregon: Association for Computational Linguistics, June.

Anderson, Margo, and Stephen E. Fienberg. 1999. "To Sample or Not to Sample? The 2000 Census Controversy." *The Journal of Interdisciplinary History* 30, no. 1 (July): 1–36. ISSN: 0022-1953, 1530-9169. https://doi.org/10.1162/002219599551895.

Augenstein, Isabelle, Tim Rocktäschel, Andreas Vlachos, and Kalina Bontcheva. 2016. "Stance Detection with Bidirectional Conditional Encoding." In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing.* Association for Computational Linguistics. https://doi.org/10.18653/v1/D16-1084.

Bakhurst, David. 2013. "Learning from Others." *Journal of Philosophy of Education* 47 (2): 187–203. ISSN: 1467-9752. https://doi.org/10.1111/1467-9752.12020.

Bauman, Richard, and Charles L. Briggs. 1990. “Poetics and Performance as Critical Perspectives on Language and Social Life.” *Annual Review of Anthropology* 19:59–88. ISSN: 0084-6570. JSTOR: 2155959.

Beach, Richard, and Chris M. Anson. 1992. “Stance and Intertextuality in Written Discourse.” *Linguistics and Education* 4, nos. 3-4 (January): 335–357. ISSN: 08985898. https://doi.org/10.1016/0898-5898(92)90007-J.

Becker, Howard Saul. 2017. *Evidence.* Chicago ; London: The University of Chicago Press. ISBN: 978-0-226-46623-1 978-0-226-46637-8.

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. “On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?” In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency,* 610–623. FAccT ’21. New York, NY, USA: Association for Computing Machinery, March 3, 2021. ISBN: 978-1-4503-8309-7. https://doi.org/10.1145/3442188.3445922.

Biber, Douglas. 1992. “The Multi-Dimensional Approach to Linguistic Analyses of Genre Variation: An Overview of Methodology and Findings.” *Computers and the Humanities* 26, nos. 5-6 (December): 331–345. ISSN: 0010-4817, 1572-8412. https://doi.org/10.1007/BF00136979.

———. 2006. “Stance in Spoken and Written University Registers.” *Journal of English for Academic Purposes* 5, no. 2 (April): 97–116. ISSN: 14751585. https://doi.org/10.1016/j.jeap.2006.05.001.

Biber, Douglas, and Edward Finegan. 1988. “Adverbial Stance Types in English.” *Discourse Processes* 11, no. 1 (January): 1–34. ISSN: 0163-853X, 1532-6950. https://doi.org/10.1080/01638538809544689.

———. 1989. “Styles of Stance in English: Lexical and Grammatical Marking of Evidentiality and Affect.” *Text - Interdisciplinary Journal for the Study of Discourse* 9, no. 1 (January 1, 1989): 93–124. ISSN: 1860-7349. https://doi.org/10.1515/text.1.1989.9.1.93.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad, and Edward Finegan. 1999. *Longman Grammar of Spoken and Written English.* Longman.

Blondel, Vincent D., Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. “Fast Unfolding of Communities in Large Networks.” *Journal of Statistical Mechanics: Theory and Experiment* 2008, no. 10 (October 9, 2008): P10008. ISSN: 1742-5468. https://doi.org/10.1088/1742-5468/2008/10/P10008. arXiv: 0803.0476.

Bonilla, Yarimar, and Jonathan Rosa. 2015. “#Ferguson: Digital Protest, Hashtag Ethnography, and the Racial Politics of Social Media in the United States.” *American Ethnologist* 42 (1): 4–17. https://doi.org/10.1111/amet.12112.

Bowman, Samuel R., Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. “A Large Annotated Corpus for Learning Natural Language Inference.” In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing,* 632–642. Lisbon, Portugal: Association for Computational Linguistics, September. https://doi.org/10.18653/v1/D15-1075.

Bras, Ronan Le, Swabha Swayamdipta, Chandra Bhagavatula, Rowan Zellers, Matthew E. Peters, Ashish Sabharwal, and Yejin Choi. 2020. “Adversarial Filters of Dataset Biases.” In *Proceedings of the 37 Th International Conference on Machine Learning.* Vienna, Austria, July 10, 2020. arXiv: 2002.04108.

Callon, Michel. 1984. “Some Elements of a Sociology of Translation: Domestication of the Scallops and the Fishermen of St Brieuc Bay.” *The Sociological Review* 32 (1_suppl 1984): 196–233. ISSN: 0038-0261. https://doi.org/10.1111/j.1467-954X.1984.tb00113.x.

Campello, Ricardo J. G. B., Davoud Moulavi, and Joerg Sander. 2013. "Density-Based Clustering Based on Hierarchical Density Estimates." In *Advances in Knowledge Discovery and Data Mining,* edited by Jian Pei, Vincent S. Tseng, Longbing Cao, Hiroshi Motoda, and Guandong Xu, redacted by David Hutchison, Takeo Kanade, Josef Kittler, Jon M. Kleinberg, Friedemann Mattern, John C. Mitchell, Moni Naor, et al., 7819:160–172. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer. ISBN: 978-3-642-37455-5 978-3-642-37456-2. https://doi.org/10.1007/978-3-642-37456-2_14.

Carson, John. 2007. *The Measure of Merit: Talents, Intelligence, and Inequality in the French and American Republics, 1750-1940.* Princeton: Princeton University Press. ISBN: 978-0-691-01715-0.

Celle, Agnès. 2009. "Hearsay Adverbs and Modality." In *Modality in English,* edited by P. Busuttil, R. Salkie, and J. van der Auwera, 269–293. Mouton de Gruyter.

Chafe, Wallace L., and Johanna Nichols, eds. 1986. *Evidentiality: The Linguistic Coding of Epistemology.* Advances in Discourse Processes 20. Norwood, N.J: Ablex Publishing Corporation. ISBN: 978-0-89391-203-1.

Chindamo, Massimo, Jens Allwood, and Elisabeth Ahlsen. 2012. "Some Suggestions for the Study of Stance in Communication." In *Proceedings of the 2012 ASE/IEEE International Conference on Social Computing and 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust,* 617–622. Amsterdam, Netherlands: IEEE, September. ISBN: 978-1-4673-5638-1 978-0-7695-4848-7. https://doi.org/10.1109/SocialCom-PASSAT.2012.89.

Choo, Esther K., Megan L. Ranney, Teresa M. Chan, N. Seth Trueger, Amy E. Walsh, Ken Tegtmeyer, Shannon O. McNamara, Ricky Y. Choi, and Christopher L. Carroll. 2015. "Twitter as a Tool for Communication and Knowledge Exchange in Academic Medicine: A Guide for Skeptics and Novices." *Medical Teacher* 37, no. 5 (May 4, 2015): 411–416. ISSN: 0142-159X. https://doi.org/10.3109/0142159X.2014.993371. pmid: 25523012.

Clift, Rebecca. 2006. "Indexing Stance: Reported Speech as an Interactional Evidential." *Journal of Sociolinguistics* 10, no. 5 (November): 569–595. ISSN: 1360-6441, 1467-9841. https://doi.org/10.1111/j.1467-9841.2006.00296.x.

Cooper, Helene, Ralph Blumenthal, and Leslie Kean. 2017. "Glowing Auras and 'Black Money': The Pentagon's Mysterious U.F.O. Program." *The New York Times: U.S.,* December 16, 2017. ISSN: 0362-4331. https://www.nytimes.com/2017/12/16/us/politics/pentagon-program-ufo-harry-reid.html.

Couldry, Nick, and Ulises A. Mejias. 2019. "Making Data Colonialism Liveable: How Might Data's Social Order Be Regulated?" *Internet Policy Review* 8, no. 2 (June 30, 2019). ISSN: 2197-6775. https://doi.org/10.14763/2019.2.1411.

D'Ignazio, Catherine, and Lauren F Klein. 2020. "The Numbers Don't Speak for Themselves." In *Data Feminism.* The MIT Press.

Daston, Lorraine. 1992. "Objectivity and the Escape from Perspective." *Social Studies of Science* 22, no. 4 (November 1, 1992): 597–618. ISSN: 0306-3127. https://doi.org/10.1177/030631292022004002.

Desrosières, Alain. 2001. "How Real Are Statistics? Four Posssible Attitudes." *Social Research* 68 (2): 339–355. ISSN: 0037-783X. JSTOR: 40971461.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." May 24, 2019. arXiv: 1810.04805 [cs].

Dey, K., Ritvik Shrivastava, and Saroj Kaushik. 2018. "Topical Stance Detection for Twitter: A Two-Phase LSTM Model Using Attention." *ECIR,* https://doi.org/10.1007/978-3-319-76941-7_40.

Didier, Emmanuel. 2002. "Sampling and Democracy: Representativeness in the First United States Surveys." *Science in Context* 15, no. 3 (September): 427–445. ISSN: 02698897, 14740664. https://doi.org/10.1017/S0269889702000558.

Dietrich, Alex. *'Just disappeared': Veteran combat pilot describes UFO sighting.* https://www.cnn.com/videos/politics/2021/05/19/alex-dietrich-ufo-sighting-ac360-intv-vpx.cnn.

Director of National Intelligence, Office of the. 2021. *Preliminary Assessment: Unidentified Aerial Phenomena.* Office of the Director of National Intelligence, June 25, 2021.

Du Bois, John W. 2007. "The Stance Triangle." In *Pragmatics & Beyond New Series,* edited by Robert Englebretson, 164:139–182. Amsterdam: John Benjamins Publishing Company. ISBN: 978-90-272-5408-5 978-90-272-9192-9. https://doi.org/10.1075/pbns.164.07du.

EcoSexuality. *Tweet.* https://twitter.com/EcoSexuality/status/13976592155656233 0.

Eghigian, Greg. 2017. "Making UFOs Make Sense: Ufology, Science, and the History of Their Mutual Mistrust." *Public Understanding of Science* 26, no. 5 (July): 612–626. ISSN: 0963-6625, 1361-6609. https://doi.org/10.1177/0963662515617706.

Espeland, Wendy Nelson, and Mitchell L. Stevens. 2008. "A Sociology of Quantification." *European Journal of Sociology* 49, no. 3 (December): 401–436. ISSN: 0003-9756, 1474-0583. https://doi.org/10.1017/S0003975609000150.

Espírito Santo, Diana, and Alejandra Vergara. 2020. “The Possible and the Impossible: Reflections on Evidence in Chilean Ufology.” *Antípoda. Revista de Antropología y Arqueología,* no. 41 (October): 125–146. ISSN: 1900-5407, 2011-4273. https://doi.org/10.7440/antipoda41.2020.06.

Feldman, Martha S., and James G. March. 1981. “Information in Organizations as Signal and Symbol.” *Administrative Science Quarterly* 26, no. 2 (June): 171. ISSN: 00018392. https://doi.org/10.2307/2392467. JSTOR: 2392467.

Ferreira, William, and Andreas Vlachos. 2016. “Emergent: A Novel Data-Set for Stance Classification.” In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* 1163–1168. San Diego, California: Association for Computational Linguistics. https://doi.org/10.18653/v1/N16-1138.

Fox, Barbara A. 2001. “Evidentiality: Authority, Responsibility, and Entitlement in English Conversation.” *Journal of Linguistic Anthropology* 11 (2): 167–192. ISSN: 10551360, 15481395. https://doi.org/10.1525/jlin.2001.11.2.167.

Fraisier, Ophélie, Guillaume Cabanac, Yoann Pitarch, Romaric Besançon, and Mohand Boughanem. 2018. “Stance Classification through Proximity-based Community Detection.” In *Proceedings of the 29th on Hypertext and Social Media,* 220–228. HT ’18. New York, NY, USA: Association for Computing Machinery, July 3, 2018. ISBN: 978-1-4503-5427-1. https://doi.org/10.1145/3209542.3209549.

Gitelman, Lisa, ed. 2013. *"Raw Data" Is an Oxymoron.* Infrastructures Series. Cambridge, Mass.: MIT Press. ISBN: 978-0-262-31232-5 978-0-262-51828-4.

Goffman, Erving. 1981. “Footing.” In *Forms of Talk.* University of Pennsylvania Publications in Conduct and Communication. Philadelphia: University of Pennsylvania Press. ISBN: 978-0-8122-7790-6 978-0-8122-1112-2.

Goodwin, Charles. 1994. “Professional Vision.” *American Anthropologist* 96 (3): 606–633. ISSN: 0002-7294. JSTOR: 682303.

Goodwin, Charles, and Marjorie Harness Goodwin. 2005. “Participation.” In *A Companion to Linguistic Anthropology,* 1st ed., edited by Alessandro Duranti, 222–244. Wiley, January. ISBN: 978-0-631-22352-8 978-0-470-99652-2. https://doi.org/10.1002/9780470996522.ch10.

Grootendorst, Maarten. 2020. “Class-Based TF-IDF.” Towards Data Science, October 19, 2020. https://towardsdatascience.com/creating-a-class-based-tf-idf-with-scikit-learn-caea7b15b858.

Hacohen-Kerner, Yaakov, Ziv Ido, and Ronen Ya’akobov. 2017. “Stance Classification of Tweets Using Skip Char Ngrams.” In *Proceedings of the 2017 European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases.* Skopje, Macedonia. https://doi.org/10.1007/978-3-319-71273-4_22.

Hanks, Michele. 2016. “Between Electricity and Spirit: Paranormal Investigation and the Creation of Doubt in England: Between Electricity and Spirit.” *American Anthropologist* 118, no. 4 (December): 811–823. ISSN: 00027294. https://doi.org/10.1111/aman.12684.

Hasan, Kazi Saidul, and Vincent Ng. 2013. “Stance Classification of Ideological Debates: Data, Models, Features, and Constraints.” In *Proceedings of the Sixth International Joint Conference on Natural Language Processing,* 1348–1356. IJCNLP 2013. Nagoya, Japan: Asian Federation of Natural Language Processing, October.

Hermans, Alexander, Lucas Beyer, and Bastian Leibe. 2017. “In Defense of the Triplet Loss for Person Re-Identification.” November 21, 2017. arXiv: 1703.07737 [cs].

Huddleston, Rodney D., and Geoffrey K. Pullum. 2002. *The Cambridge Grammar of the English Language.* Cambridge, UK ; New York: Cambridge University Press. ISBN: 978-0-521-43146-0.

Hunston, Susan, and Geoff Thompson, eds. 2000. *Evaluation in Text: Authorial Stance and the Construction of Discourse.* Oxford Linguistics. Oxford ; New York: Oxford University Press. ISBN: 978-0-19-823854-6.

Hwang, Tim, and Karen Levy. 2015. "'The Cloud' and Other Dangerous Metaphors." *The Atlantic,* January 20, 2015, 7:07 p.m. (Z). https://www.theatlantic.com/technology/archive/2015/01/the-cloud-and-other-dangerous-metaphors/384518/.

Hyland, Ken. 1996. "Writing Without Conviction: Hedging in Science Research Articles." *Applied Linguistics* 17, no. 4 (December 1, 1996): 433–454. ISSN: 0142-6001, 1477-450X. https://doi.org/10.1093/applin/17.4.433.

———. 2002. "Authority and Invisibility." *Journal of Pragmatics* 34, no. 8 (August): 1091–1112. ISSN: 03782166. https://doi.org/10.1016/S0378-2166(02)00035-8.

IndiaJenkins1. *Tweet.* https://twitter.com/IndiaJenkins1/status/1397666004457639940.

Jaffe, Alexandra. 2009. *Stance: Sociolinguistic Perspectives.* Oxford University Press, June 4, 2009. ISBN: 978-0-19-533164-6. https://doi.org/10.1093/acprof:oso/9780195331646.001.0001.

Jakobson, Roman, Linda R. Waugh, and Monique Monville-Burston. 1990. *On Language.* Cambridge, Mass: Harvard University Press. ISBN: 978-0-674-63535-7.

Jones, Edward E, and Victor A Harris. 1967. "The Attribution of Attitudes." *Journal of Experimental Social Psychology* 3, no. 1 (January): 1–24. ISSN: 00221031. https://doi.org/10.1016/0022-1031(67)90034-0.

Jones, Graham M. 2014. “Secrecy.” *Annual Review of Anthropology* 43 (1): 53–69. https://doi.org/10.1146/annurev-anthro-102313-030058.

Jurafsky, Dan, Elizabeth Shriberg, Barbara A. Fox, and T. Curl. 1998. “Lexical, Prosodic, and Syntactic Cues for Dialog Acts.” In *Workshop On Discourse Relations And Discourse Markers.*

Kalthoff, Herbert. 2005. “Practices of Calculation: Economic Representations and Risk Management.” *Theory, Culture & Society* 22, no. 2 (April): 69–97. ISSN: 0263-2764, 1460-3616. https://doi.org/10.1177/0263276405051666.

Kärkkäinen, Elise. 2003. *Epistemic Stance in English Conversation: A Description of Its Interactional Functions, with a Focus on I Think.* Pragmatics and Beyond, N.S., 115. Amsterdam: Benjamins. ISBN: 978-90-272-5357-6 978-1-58811-444-0.

Keyes, Os. 2018. “The Misgendering Machines: Trans/HCI Implications of Automatic Gender Recognition.” *Proceedings of the ACM on Human-Computer Interaction* 2 (CSCW): 1–22.

Kiesling, Scott F. 2022. “Stance and Stancetaking.” *Annual Review of Linguistics* 8, no. 1 (January 14, 2022): 409–426. ISSN: 2333-9683, 2333-9691. https://doi.org/10.1146/annurev-linguistics-031120-121256.

Kiesling, Scott F., Umashanthi Pavalanathan, Jim Fitzpatrick, Xiaochuang Han, and Jacob Eisenstein. 2018. “Interactional Stancetaking in Online Forums.” *Computational Linguistics* 44, no. 4 (December): 683–718. ISSN: 0891-2017, 1530-9312. https://doi.org/10.1162/coli_a_00334.

KitCMiller2. *Tweet.* https://twitter.com/KitCMiller2/status/1397659452971667459.

Kockelman, Paul. 2004. “Stance and Subjectivity.” *Journal of Linguistic Anthropology* 14, no. 2 (December): 127–150. ISSN: 1055-1360, 1548-1395. https://doi.org/10.1525/jlin.2004.14.2.127.

Kube, Courtney, and Adam Edelman. 2021. "UFO Report: Government Can't Explain 143 of 144 Mysterious Flying Objects, Blames Limited Data." NBC News, June 25, 2021. https://www.nbcnews.com/politics/politics-news/ufo-report-government-can-t-explain-143-144-mysterious-flying-n1272390.

Küçük, Dilek, and Fazli Can. 2021. "Stance Detection: A Survey." *ACM Computing Surveys* 53, no. 1 (January 31, 2021): 1–37. ISSN: 0360-0300, 1557-7341. https://doi.org/10.1145/3369026.

KyleParanormal and Pagan. 2021. *Interview: AP Strange on #ufotwitter and the Weird!* Chaos and Shadow.

larryca66028461. *Tweet.* https://twitter.com/larryca66028461/status/1397658976838467594.

Latour, Bruno. 1988. *Science in Action: How to Follow Scientists and Engineers through Society.* Cambridge, Mass: Harvard Univ. Press. ISBN: 978-0-674-79291-3.

Lee, Crystal, Tanya Yang, Gabrielle D Inchoco, Graham M. Jones, and Arvind Satyanarayan. 2021. "Viral Visualizations: How Coronavirus Skeptics Use Orthodox Data Practices to Promote Unorthodox Science Online." In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems,* 1–18. CHI '21: CHI Conference on Human Factors in Computing Systems. Yokohama Japan: ACM, May 6, 2021. ISBN: 978-1-4503-8096-6. https://doi.org/10.1145/3411764.3445211.

Leite, Christopher C., and Can E. Mutlu. 2017. "The Social Life of Data: The Production of Political Facts in EU Policy Governance." *Global Governance* (Leiden, Netherlands Antilles) 23, no. 1 (January–March): 71–82. ISSN: 10752846. https://doi.org/http://dx.doi.org.libproxy.mit.edu/10.1163/19426720-02301007.

Lepselter, Susan. 2016. *The Resonance of Unseen Things: Poetics, Power, Captivity, and UFOs in the American Uncanny.* University of Michigan Press. ISBN: 978-0-472-90065-7 978-0-472-07294-1, accessed May 3, 2022. http://www.doabooks.org/doab?func=fulltext&uiLanguage=en&rid=19057.

Lewis-Kraus, Gideon. 2021. "How the Pentagon Started Taking U.F.O.s Seriously." *The New Yorker,* May 10, 2021.

Ling, Wang, Chris Dyer, Alan W Black, and Isabel Trancoso. 2015. "Two/Too Simple Adaptations of Word2Vec for Syntax Problems." In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* 1299–1304. Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Denver, Colorado: Association for Computational Linguistics. https://doi.org/10.3115/v1/N15-1142.

live4literacy. *Tweet.* https://twitter.com/live4literacy/status/1397686711455518723.

Maaten, Laurens van der, and Geoffrey Hinton. 2008. "Visualizing Data Using T-SNE." *Journal of Machine Learning Research* 9 (86): 2579–2605. ISSN: 1533-7928.

Madmom42004988. *Tweet.* https://twitter.com/Madmom42004988/status/1397695239729582086.

Mandavilli, Apoorva. 2011. "Peer Review: Trial By Twitter." *Nature* 469:286–287.

Martin, J. R., and P. R. R. White. 2005. *The Language of Evaluation: Appraisal in English.* New York: Palgrave Macmillan. ISBN: 978-1-4039-0409-6.

Martin, J.R. 2000. "Beyond Exchange: Appraisal Systems in English." In *Evaluation in Text: Authorial Stance and the Construction of Discourse,* edited by S. Hunston and G. Thompson, 142–175. Oxford University Press.

Martini, Francesco. 2017. "Hearsay Viewed through the Lens of Trust, Reputation and Coherence." *Synthese* 194, no. 10 (October): 4083–4099. ISSN: 0039-7857, 1573-0964. https://doi.org/10.1007/s11229-016-1128-7.

McDowell, John Henry. 1998. *Mind, Value, and Reality.* Cambridge, Mass: Harvard University Press. ISBN: 978-0-674-57613-1.

McInnes, Leland, John Healy, and James Melville. 2020. "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction." September 17, 2020. arXiv: 1802.03426 [cs, stat].

Mejias, Ulises A., and Nick Couldry. 2019. "Datafication." *Internet Policy Review* 8, no. 4 (November 29, 2019). ISSN: 2197-6775. https://doi.org/10.14763/2019.4.1428.

Mithun, Marianne. 1986. "Evidential Diachrony in Northern Iroquoian." In *Evidentiality: The Linguistic Coding of Epistemology,* edited by Wallace Chafe and Johanna Nichols. Norwood, New Jersey: Ablex Publishing Corporation.

Mohammad, Saif, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. "SemEval-2016 Task 6: Detecting Stance in Tweets." In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016),* 31–41. San Diego, California: Association for Computational Linguistics, June. https://doi.org/10.18653/v1/S16-1003.

Mohammad, Saif M., Parinaz Sobhani, and Svetlana Kiritchenko. 2017. "Stance and Sentiment in Tweets." *ACM Transactions on Internet Technology* 17, no. 3 (July 14, 2017): 1–23. ISSN: 1533-5399, 1557-6051. https://doi.org/10.1145/3003433.

Mushin, Ilana. 2001. *Evidentiality and Epistemological Stance: Narrative Retelling.* Pragmatics & Beyond, new ser. 87. Amsterdam ; Philadelphia: John Benjamins Pub. Co. ISBN: 978-1-58811-033-6 978-90-272-5106-0.

Newman, M. E. J., and M. Girvan. 2004. “Finding and Evaluating Community Structure in Networks.” *Physical Review E* 69, no. 2 (February 26, 2004): 026113. ISSN: 1539-3755, 1550-2376. https://doi.org/10.1103/PhysRevE.69.026113. arXiv: cond-mat/0308217.

Niven, Timothy, and Hung-Yu Kao. 2019. “Probing Neural Network Comprehension of Natural Language Arguments.” In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics,* 4658–4664. Florence, Italy: Association for Computational Linguistics, July. https://doi.org/10.18653/v1/P19-1459.

Nussbaum, Martha C. 1986. *The Fragility of Goodness: Luck and Ethics in Greek Tragedy and Philosophy.* Cambridge, U.K. ; New York: Cambridge University Press.

PaintingSurfer. *Tweet.* https://twitter.com/PaintingSurfer/status/1405035558414348289.

Palmer, F. R. 1986. *Mood and Modality.* Cambridge, UK; New York: Cambridge University Press. ISBN: 978-1-139-16717-8.

Pasulka, Diana Walsh. 2019. *American Cosmic: UFOs, Religions, Techonology.* New York, NY: Oxford University Press. ISBN: 978-0-19-069349-7 978-0-19-069350-3.

Pavalanathan, Umashanthi, Jim Fitzpatrick, Scott Kiesling, and Jacob Eisenstein. 2017. “A Multidimensional Lexicon for Interpersonal Stancetaking.” In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics,* 884–895. Vancouver, Canada: Association for Computational Linguistics. https://doi.org/10.18653/v1/P17-1082.

Phenomenon, Engaging the, dir. 2020. *Welcome To UFO Twitter (#ufotwitter).* March 12, 2020.

Pomerleau, Dean, and Delip Rao. 2017. *Fake News Detection Challenge: FNC-1.* http://www.fakenewschallenge.org/.

Porter, Theodore M. 1995. *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life.* Princeton University Press, January 1, 1995. ISBN: 978-0-691-21054-4. https://doi.org/10.1515/9780691210544.

Puschmann, Cornelius, and Jean Burgess. 2014. "Big Data, Big Questions| Metaphors of Big Data." *International Journal of Communication* 8, no. 0 (0 2014): 20. ISSN: 1932-8036.

Qiu, Minghui, Yanchuan Sim, Noah A. Smith, and Jing Jiang. 2015. "Modeling User Arguments, Interactions and Attributes for Stance Prediction in Online Debate Forums." *Proceedings of the 2015 SIAM International Conference on Data Mining: April 30 - May 2, Vancouver, Canada* (May 2, 2015): 855–863. https://doi.org/10.1137/1.9781611974010.96.

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. 1985. *A Comprehensive Grammar of the English Language.* London ; New York: Longman. ISBN: 978-0-582-51734-9.

Reimers, Nils, and Iryna Gurevych. 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),* 3980–3990. Hong Kong, China: Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1410.

Rozumko, Agata. 2019. "Between Acknowledgement and Countering: Interpersonal Functions of English Reportative Adverbs." *Journal of Pragmatics* 140 (January): 1–11. ISSN: 03782166. https://doi.org/10.1016/j.pragma.2018.11.011.

Runge, Kristin K., Sara K. Yeo, Michael Cacciatore, Dietram A. Scheufele, Dominique Brossard, Michael Xenos, Ashley Anderson, et al. 2013. “Tweeting Nano: How Public Discourses about Nanotechnology Develop in Social Media Environments.” *Journal of Nanoparticle Research* 15, no. 1 (January): 1381. ISSN: 1388-0764, 1572-896X. https://doi.org/10.1007/s11051-012-1381-8.

Sadowski, Jathan. 2019. “When Data Is Capital: Datafication, Accumulation, and Extraction.” *Big Data & Society* 6, no. 1 (January 1, 2019): 2053951718820549. ISSN: 2053-9517. https://doi.org/10.1177/2053951718820549.

San Roque, Lila. 2019. “Evidentiality.” *Annual Review of Anthropology.*

San Roque, Lila, Simeon Floyd, and Elisabeth Norcliffe. 2018. “Egophoricity: An Introduction.” In *Typological Studies in Language,* edited by Simeon Floyd, Elisabeth Norcliffe, and Lila San Roque, 118:1–78. Amsterdam: John Benjamins Publishing Company, April 16, 2018. ISBN: 978-90-272-0699-2 978-90-272-6554-8. https://doi.org/10.1075/tsl.118.01san.

Schwartz, Leo. 2022. “Amateur Open-Source Researchers Went Viral Unpacking the War in Ukraine.” *Rest of World,* March 7, 2022. https://restofworld.org/2022/osint-viral-ukraine/.

Sidnell, Jack. 2012. "“Who Knows Best?”: Evidentiality and Epistemic Asymmetry in Conversation.” *Pragmatics and Society* 3, no. 2 (October 23, 2012): 294–320. ISSN: 1878-9714, 1878-9722. https://doi.org/10.1075/ps.3.2.08sid.

Simaki, Vasiliki, Panagiotis Simakis, Carita Paradis, and Andreas Kerren. 2018. “Detection of Stance-Related Characteristics in Social Media Text.” In *Proceedings of the 10th Hellenic Conference on Artificial Intelligence,* 1–7. Patras Greece: Association for Computing Machinery, July 9, 2018. ISBN: 978-1-4503-6433-1. https://doi.org/10.1145/3200947.3201017.

Sobhani, Parinaz, Diana Inkpen, and Xiaodan Zhu. 2017. "A Dataset for Multi-Target Stance Detection." In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics,* 551–557. Valencia, Spain: Association for Computational Linguistics. https://doi.org/10.18653/v1/E17-2088.

Stevens, Nikki, and Os Keyes. 2021. "Seeing Infrastructure: Race, Facial Recognition and the Politics of Data." *Cultural Studies* 35, nos. 4-5 (September 3, 2021): 833–853. ISSN: 0950-2386. https://doi.org/10.1080/09502386.2021.1895252.

Thompson, G., and Y. Ye. 1991. "Evaluation in the Reporting Verbs Used in Academic Papers." *Applied Linguistics* 12, no. 4 (December 1, 1991): 365–382. ISSN: 0142-6001, 1477-450X. https://doi.org/10.1093/applin/12.4.365.

Tsakalidis, Adam, Nikolaos Aletras, A. Cristea, and Maria Liakata. 2018. "Nowcasting the Stance of Social Media Users in a Sudden Vote: The Case of the Greek Referendum." In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management.* New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3269206.3271783.

Vick, Karl. 2022. "Bellingcat's Eliot Higgins Explains Why Ukraine Is Winning the Information War." *Time,* March 9, 2022. https://time.com/6155869/bellingcat-eliot-higgins-ukraine-open-source-intelligence/.

Viljoen, Salomé. 2021. "A Relational Theory of Data Governance." *Yale Law Journal* 131.

Weber, Max. 1978. *Economy and Society.* Berkeley: University of California Press. ISBN: 978-0-520-28002-1.

Wei, Penghui, Junjie Lin, and W. Mao. 2018. "Multi-Target Stance Detection via a Dynamic Memory-Augmented Network." *SIGIR,* https://doi.org/10.1145/3209978.3210145.

Wei, Wan, Xiao Zhang, Xuqin Liu, Wei Chen, and Tengjiao Wang. 2016. "Pkudblab at SemEval-2016 Task 6 : A Specific Convolutional Neural Network System for Effective Stance Detection." In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016),* 384–388. San Diego, California: Association for Computational Linguistics. https://doi.org/10.18653/v1/S16-1062.

Wierzbicka, Anna. 2006. *English: Meaning and Culture.* Oxford ; New York: Oxford University Press. ISBN: 978-0-19-517474-8 978-0-19-517475-5.

Willett, Thomas. 1988. "A Cross-Linguistic Survey of the Grammaticization of Evidentiality." *Studies in Language* 12, no. 1 (January 1, 1988): 51–97. ISSN: 0378-4177, 1569-9978. https://doi.org/10.1075/sl.12.1.04wil.

Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, et al. 2020. "Transformers: State-of-the-Art Natural Language Processing." In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations,* 38–45. Online: Association for Computational Linguistics, October. https://doi.org/10.18653/v1/2020.emnlp-demos.6.

Zhang, Shaodian, Lin Qiu, Frank Chen, Weinan Zhang, Yong Yu, and Noémie Elhadad. 2017. "We Make Choices We Think Are Going to Save Us: Debate and Stance Identification for Online Breast Cancer CAM Discussions." In *Proceedings of the 26th International Conference on World Wide Web Companion,* 1073–1081. Perth, Australia: Association for Computing Machinery. ISBN: 978-1-4503-4914-7. https://doi.org/10.1145/3041021.3055134.

Zhou, Yiwei, A. Cristea, and Lei Shi. 2017. "Connecting Targets to Tweets: Semantic Attention-Based Model for Target-Specific Stance Detection." In *Proceedings of the 18th International Conference on Web Information Systems Engineering.* https://doi.org/10.1007/978-3-319-68783-4_2.

Zubiaga, Arkaitz, Ahmet Aker, Kalina Bontcheva, Maria Liakata, and Rob Procter. 2018. “Detection and Resolution of Rumours in Social Media: A Survey.” *ACM Computing Surveys* 51, no. 2 (June 2, 2018): 1–36. ISSN: 0360-0300, 1557-7341. https://doi.org/10.1145/3161603. arXiv: 1704.00656.