ZEITSCHRIFT FÜR SPRACHWISSENSCHAFT
Journal of the Linguistic Society of Germany

Volume 44, Issue 1, Article 3
https://doi.org/10.18148/zs/2025-2003

Maite Taboada, Simon Fraser University, Canada, mtaboada@sfu.ca

Computational evaluation

Representing and processing evaluative language

Abstract: I propose that *computational evaluation* is an emerging field of research, one that applies computational techniques to the representation and processing of evaluative language, associating evaluative meanings with expressions of human language. The study of evaluative language has a long history in linguistics, encompassing research on attitude, subjectivity, point of view, and evidentiality, with more recent studies on appraisal or emotion language. At the same time, computational linguistics has by now accumulated a back catalogue of research going back a couple of decades into how we can extract evaluation, sentiment, and opinion automatically from text. I briefly survey this history, to then outline a proposal that the study of evaluative language from a computational point of view crosscuts all levels of language, from morphology and the lexicon to figuration, and requires a comprehensive understanding of language. By way of illustration, I will discuss research on appraisal, abusive language online, and the use of metaphors in the expression of negative opinion. This work has applications in content moderation, detection of misinformation, or information retrieval, but it is also interesting in its own right, as a theoretical field in linguistics and computational linguistics.

Keywords: evaluative language, subjectivity, figures of speech, metaphor, corpus linguistics, computational linguistics

This invited contribution is based on the plenary lecture given on **28 February 2024** at the Annual Conference of the German Linguistic Society.

Cite as: Taboada, Maite. 2025. Computational evaluation: Representing and processing evaluative language. *Zeitschrift für Sprachwissenschaft* 44(1). https://doi.org/10.18148/zs/2025-2003.

1 Introduction: computational evaluation as a subfield of linguistics

Research on evaluative and subjective language has been a regular preoccupation of linguists, with many and varied approaches to how language conveys subjectivity, emotion, opinion, or evidentiality. At the same time, research in computational linguistics (CL) and natural language processing (NLP) has focused on how to extract and label evaluative meaning in text, especially in the field of sentiment analysis. My proposal is that, relatedly but independently of sentiment analysis, linguists have important contributions to make in the subfield of computational evaluation, one where the goal is to represent and process evaluative language, associating evaluative meanings with expressions of human language.

My goal in this paper is to advance the notion of computational evaluation as an overarching approach to the study and computational processing of language. I will illustrate this with projects on evaluative language that we have carried out in my lab in the last few years, stressing how we have needed multiple levels of language to approach them. I will note that I discuss text analysis exclusively, including analysis of speech that has been transcribed as text. Detection of emotion and evaluation from speech and sign language requires additional techniques and considerations.

The study of evaluative language has a long history in linguistics, encompassing research on affect, subjectivity, point of view, and evidentiality, with more recent studies on appraisal or emotion language (Chafe & Nichols 1986; Biber & Finegan 1988; Giannakidou 1995; Wierzbicka 1999; Martin & White 2005; Bednarek 2008; Hunston 2011; Dancygier & Sweetser 2012; Gutzmann 2019). What I want to draw from all this research is that, in linguistics, we understand the complexities of evaluation and attitude and we can use that extensive background on how evaluative language works to "computationalize" those insights.

I like to frame the concept of computational evaluation in terms of the difference between CL and NLP. Whether there is a difference, and what the difference is, has long been a debate in this and related fields (Raskin 1987; Tsujii 2021). There are many ways of looking at the division of labour, but one that I found particularly helpful is Josef Fruehwald's formulation (Fruehwald 2024). He proposes that CL is much broader than NLP and concerned not just with applications. He suggests that a good way to approach the distinction is to think about the set of things x that are part of CL and not part of NLP, as shown in (1), from Fruehwald (2024).

(1) $\{ x \mid x \in CL \text{ and } x \neq NLP \}$

Some of Fruehwald's examples for elements in that set include parsers for minimalist syntax (Berwick & Stabler 2019), where the goal is to model the minimalist framework and not necessarily to build a good or efficient parser, as we know that the fastest and more accurate parsers are in the dependency formalism (De Marneffe et al. 2021; Jurafsky & Martin 2024). Building a fast and efficient parser is an NLP problem; building a parser to test a linguistic theory is a CL problem. In that set of CL problems are also agent-based models, such as models of how language change spreads through a community (Stanford & Kenny 2013).

Inspired by this approach, my question is what is the set of *x* that belongs in computational evaluation. An analogous formula is in (2).

(2) $\{ x \mid x \in \textit{Computational Evaluation and } x \neq \textit{NLP} \}$

I define computational evaluation as an approach to understanding and representing evaluative language in a computational way. It is perhaps most straightforward to define it as the range of studies that one can undertake and the set of questions that one can ask which would fall under the rubric of computational evaluation. Some examples are in the realm of sentiment analysis, which is fundamentally a part of NLP. There are, however, aspects of sentiment analysis that NLP may not be interested in, because they do not directly address the problem of how best to determine the sentiment of a text. Such aspects may concern the general linguistic patterns for sentiment, including patterns that are difficult to extract automatically, because they are heavily context dependent. Another aspect may be the difference in how positive and negative evaluation is expressed, and what linguistic devices negative evaluation employs more frequently, one of the issues I explore, especially as concerns how figurative language is used to convey negative evaluation. Such studies are of interest to linguists and perhaps to researchers in computational evaluation, but they would not necessarily be of much interest to NLP researchers, as they may not directly contribute to better sentiment analysis systems.

Another aspirational example could be an approach that formalizes how people evaluate something collaboratively, in conversation, or how they contribute to a debate in a constructive way. I say this is aspirational because the studies I have conducted so far have only scratched the surface of how we could formalize collaborative evaluation (Canute et al. 2023; Kolhatkar et al. 2023).

Approaching the definition in terms of how it is similar to other fields, I would propose that it is similar to computational semantics or computational phonology. Blackburn & Bos (2003: 27) state that computational semantics has as a goal to "automate the process of associating semantic representations with expressions of human language". Computational phonology, according to the Association for Computational Linguistics, is "the application of formal and computational techniques to the representation and processing of phonological information".[1] Similarly, computational evaluation has as a goal to understand and represent evaluative language, so that we can associate expressions of human language to adequate representations or abstractions which capture the meaning of the evaluation.

[1] https://aclweb.org/aclwiki/Computational_Phonology, 12 July 2024

I return to what is inside computational evaluation below. First, I would like to provide some examples of the study of evaluative language from a computational point of view that I believe are firmly grounded in NLP and are not necessarily examples of computational evaluation, so that we can appreciate the difference. I then define and illustrate how to do computational evaluation in Section 3, to then expand on one project on computational evaluation, the goal of which is to detect metaphors and other forms of figurative language that are used in abusive language online, in Section 4.

2 Evaluative language in NLP

In this section, I address research under the rubric of either NLP or CL but in either case oriented towards applications of computational processing, rather than towards understanding language from a computational point of view.

NLP has by now accumulated a large back catalogue of research going back a couple of decades into how we can extract evaluation, sentiment, and opinion automatically from text. Some of that work is more theoretical, with a focus on how we define and operationalize concepts such as subjectivity or viewpoint or how we can capture opinion and sentiment. An example is a series of papers by Janyce Wiebe and colleagues formalizing subjectivity (Wiebe 1994; Wiebe et al. 2004; Wilson et al. 2009), inspired by the work of Ann Banfield on point of view in narrative (Banfield 1982). Wiebe et al. (2004: 277) define subjectivity as "aspects of language used to express opinions, evaluations, and speculations". From that definition, the researchers embark on an ambitious research project that consists of identifying linguistic markers of subjectivity, organizing them in dictionaries, and building an NLP system to automatically identify if a text is subjective and, if so, whether positive or negative. This line of research is an excellent example of NLP work that is inspired and grounded in linguistics and has the dual goal of learning something about subjective language and of resulting in an application, an NLP system.

Some of the NLP research on evaluative language is more application oriented. Here, we can include the large field of automated content moderation (Gillespie 2020). In general terms, automated content moderation does not have as a goal the understanding or representation of the language to be moderated. Its primary goal is to produce a system that can reliable (and hopefully transparently) moderate online comments and online content. We can also include work on detecting the language of misinformation, which is often evaluative (Asr & Taboada 2019; Guo et al. 2022; Asr et al. 2023). As an illustration of *x* that fluctuates between belonging in computational evaluation or NLP, I will expand here on sentiment analysis.

Sentiment analysis, which goes by other names such as opinion mining, subjectivity analysis, or tone extraction, is, in the most basic definition, the classification of texts based on subjective content. When firmly in NLP, this is a text classification problem, akin to spam detection or fake news detection, using machine learning and AI techniques.

The definition includes classification of *texts*. Naturally, what we mean by *text* can be many things, from an entire document (a Bank of Canada statement, a UN press release, a news article, a blog post, a comment after a news article or a YouTube video), to a message in one of the many short comment platforms available at the time of writing (e.g., Instagram, Twitter/X, Threads, Mastodon, BlueSky). We could also analyze private chat messages, WhatsApp messages, messages on the company Slack channel, a sentence, a news headline. Anything that can be broadly construed as a *text* is fair game for sentiment analysis. As I have done before, I will focus here largely on text, ignoring the very important multimodal aspect of sentiment and evaluation, even when conveyed digitally and primarily in textual form.

The input in sentiment analysis is a text, whatever form that takes, and the output is a score, representing the *sentiment* of the text (Taboada 2016). The score can be on different scales, a binary positive or negative, a five-point system like that of Amazon reviews, or a completely different scale which can ultimately be reduced to *positive* or *negative*. For instance, in a study of Reddit comments, the company Aware uses the labels *toxic* and *healthy*, but ultimately, what they are doing is sentiment analysis, and both can be mapped to a positive and negative scale (Donahue et al. 2025). Examples of previous research where the meanings of positive and negative, and the texts, are slightly different, include the following:

- Good or bad news headlines (Ku et al. 2006; Balahur et al. 2010)
- Likes and dislikes in books, movies, and consumer products (Pang et al. 2002; Taboada et al. 2011)
- Pros and cons of a product (Kim & Hovy 2006)
- Candidate likely or unlikely to win an election (Kim & Hovy 2007; Mohammad et al. 2015)
- Support or opposition for proposed legislation (Bansal et al. 2008)
- Agreement or disagreement with a topic (Malouf & Mullen 2008)
- Arguments in favour or against a topic (Somasundaran & Wiebe 2009; Stab & Gurevych 2017)
- Improvement or death in medical texts (Niu et al. 2005)
- Depression or not from social media postings (Zucco et al. 2017; Babu & Kanaga 2022)
- Stock price likely to go up or down (Bouktif et al. 2020)
- Constructive or toxic comment or post (Cavasso & Taboada 2021; Kolhatkar et al. 2023)
- Good or bad employee morale (Xie et al. 2022; Donahue et al. 2025)

Given the diversity of problems sentiment analysis tries to tackle, it is remarkable that approaches typically fall into just two main camps, either knowledge-based or machine learning (Taboada 2016; Taboada et al. 2011). The first approach, the knowledge-based, lexicon-based, dictionary-based, or rule-based approach, starts with creating dictionaries of positive and negative words and various rules. For instance, there could be a rule that when a positive word is in the scope of negation, the polarity is reversed. Or that when a *sentiment* word is in the scope of an irrealis or nonveridical element (Giannakidou 2001), the strength is lowered. In Table 1, we see some words from the SO-CAL dictionary, which has polarity for each word in a scale from +5 to −5 (Taboada et al. 2011). The second column shows that polarity when it is in the scope of negation, i.e., reversed (*not good*). The third column includes examples of the word in the scope of an irrealis marker, such as *could be good* or *should have been good*, where we have applied a simple rule that brings the value of the word down by two points.

Table 1: Polarities and contextual changes for select words

Word	Polarity	With negation	With irrealis
good	3	−3	1
excellent	5	−5	3
masterpiece	5	−5	3
bad	−3	3	−1
terrible	−5	5	−3
disaster	−4	4	−2

The development of such a system consists of carefully crafting dictionaries, often tailored to the specific register that needs to be analyzed. Specific words will have different polarities depending on the context. For instance, a *loud* restaurant is usually negative. A *loud* speaker is a good thing (*loud enough for a big room*; *loud enough to use outdoors*). Rules to be applied also need to be carefully considered. We learned in our work that the polarity reversal for negation that is reflected in Table 1 above is too simple, as negation is frequently more subtle, instead acting as a shift on a scale. For example, given the word *excellent* with a value of 5 on a scale, the negated *not excellent* does not seem to convey a meaning of −5 but rather a downtoned value, perhaps a 1 or −1. The effect of negation is not a polarity switch, but a polarity shift.

The application of these systems to new data consists of extracting words from a new text and comparing them to words in the dictionary. The words are labelled with their polarity and rules are applied when contextual information may change the polarity. Eventually, all the words in the text are usually averaged to arrive at a final score for the text. The average method can be improved if we have enough knowledge of the structure of the text. We found that, for reviews, it is better to assign higher weight to words towards the end of the review, as words at the beginning tend to provide context and are not as relevant to the overall evaluation (Taboada & Grieve 2004), as we will see below.

The other main approach to sentiment analysis is a statistical method, mostly based on machine learning, on learning patterns from annotated data. To learn those patterns, researchers prepare training data, that is, texts that are varied enough but similar enough to the texts that they want to process. Each text also has a label, indicating whether the text conveys positive or negative sentiment. The labels can also be more granular, perhaps a five-star system like that of reviews on many sites. The training phase consists of applying a machine learning algorithm to learn patterns from the text, extracting relevant features. The features may be as simple as frequency or presence of positive or negative words but also length of text, length of sentences, what words appear and not, etc. For instance, in a previous study, we found that the word *plot* was predictive of a negative review for a movie (Taboada et al. 2011). It looks like people only mention plots when they have something bad to say about them. In that case, *plot* is not a negative word per se, but instead a feature of negative reviews.

In that training phase in machine learning, what we do is apply an algorithm to those features we extracted from the training data. The algorithms come in different flavours, but the important thing is that the algorithm builds a classifier. To use the system “in the wild”, we process a new text, extract features from it and ask the classifier whether those features are correlated with positive or negative reviews in the training data. Then, we obtain a label and we can assign that label to new data.

Each method has its downsides, of course. The lexicon method relies on hand-crafted dictionaries that are static; a human must add to them or curate them. The machine learning dictionary learns to predict very specific characteristics of the training data, what is known as *overfitting*. Researchers have reported that a method was correlating the name *Angelina Jolie* with negative reviews. This is because it was based on a corpus of movie reviews that contained many negative reviews of the movie *Tomb Raider 2*, from 2003, which Angelina Jolie starred in (Pang et al. 2002).

More modern methods of machine learning incorporate other forms of learning, including semi-supervised or unsupervised methods, some of them using deep learning (Hoang et al. 2019; Roccabruna et al. 2022; Cui et al. 2023; Badia et al. 2024).

In all cases, the complexity of the sentiment analysis task is derived from the inherent complexity of evaluative language, which is heavily context dependent. Most systems treat a text as a bag of words, counting occurrences of certain words or measuring whole text characteristics, like length. In addition to the phenomena I illustrated in Table 1, which are local to words and phrases, there are whole clause and whole text phenomena that are difficult to capture automatically. The first one is irony and sarcasm, a very complex linguistic phenomenon. To illustrate it, let us look at one of my favourite examples, a review of the 2000 album by rapper Shyne. The review site had a prompt, asking people to say where or when they would listen to the music. The prompt was “This is great music to play while…”. A reviewer completed the prompt with the text in (3). Of course, the word *garbage* may have been enough to classify this as negative, but the sarcasm affects the interpretation of the whole sentence.

(3) *"This is great music to play while..." getting ready to go out, because hearing this garbage makes you want to leave as fast as possible.*

One more example I would like to offer is the role of world knowledge in the interpretation of evaluation, again with an example from a corpus of reviews (Taboada 2008). The text in (4) is part of a review of a small SUV. Without any knowledge of the world, an automated system may pick up *a lot of leg room* as a positive thing for a vehicle. World knowledge comes in, of course, when we realize that children have small legs and this means that there is, in fact, not much leg room for most people. In addition to background knowledge, the text also plays with sarcasm to make a point.

(4) *Speaking of kids in the back-seat, this little truck actually has a lot of leg room... for kids*

In summary, accurate sentiment analysis requires knowledge of context and deep linguistic knowledge. I have been saying for 20 years that we need context to do sentiment analysis well. Farah Benamara, Yannick Mathieu and I wrote a 65-page paper saying that context is important and proposing a model for representing it (Benamara et al. 2017). In our model, we had an operator, Ω, to represent the intrinsic properties of a sentiment expression, and a set of functions to capture extrinsic properties, so that you can update the values in Ω. The intrinsic properties are shown in (5), namely entity, aspect, sentiment, and holder. Then, in (6), we see a set of functions that may update the properties of Ω, including aspects such as negation, sarcasm, or world knowledge. Finally, the process of updating Ω is in (7), whereby whenever we encounter a property in an update function, we can update the values in Ω to reflect that contextual information.

(5) $\Omega = (\mathrm{e}, \mathrm{a}, \mathrm{s}, \mathrm{h})$

(6) $\mathfrak{F} = \{ F_1, \ldots, F_n \}$

(7) $\forall\, F_i \in \mathfrak{F},\ F_i : \Omega \mapsto \mathit{Update}(\Omega)$

This is an example of computational evaluation, an approach that formalizes the type of information needed to understand and represent evaluative language. I should emphasize that many other scholars understood the need to address context for rigorous processing of evaluative language; some of that work is collected in a special issue of *Computational Linguistics* that I co-edited (Benamara et al. 2018).

With the advent of large language models (which I will refer to as LLMs), we have opened new avenues of exploration. One question that many researchers have explored is whether LLMs capture context, adequately or at all. This is, in a way, a reverse engineering process. In previous supervised machine learning approaches, what we had was features that we all knew and understood well and then added more features, some of them discourse and context features, like nuanced ways of capturing negation, for instance. Then, through a process of evaluation and further training, we understood better which features were helpful and which were not. This, of course, was accompanied by experimentation with different types of algorithms for the training itself.

Research has shown that language models can infer discourse structure from documents, performing as well as some human or other automated baselines (Huber & Carenini 2022). Language models are also able to capture discourse entities and their relations, in a sort of coreference chain structure (Loáiciga et al. 2022). Recent research also shows that context and LLMs can work together to enhance sentiment analysis (Ein-Dor et al. 2022). Much of this research is theoretical; there are also systems out there that claim to have solved the holy grail of sentiment analysis, using it for actual practical and consequential applications like stock market prediction (Global Markets Training 2024).

This takes us to what I would like to discuss next. I would like to focus away from chasing state of the art results and having applications that, fundamentally, we know are not capturing what is important for understanding and processing evaluative language and instead to go back to fundamentals, to representing and processing evaluative language in a principled way.

3 How to do computational evaluation

It is perhaps best to answer this question with examples, immodestly from my own work. I will discuss research in my lab on sentiment and opinion. Although some of it was partly within NLP and with applications in mind, it is still a contribution to computational evaluation, in that the primary motivation was to understand how evaluative language works. I start with three examples of research on patterns of opinion and then I continue with patterns of opinion in online news comments.

First, let us examine linguistic patterns for sentiment and three aspects that reveal the linguistic complexity of sentiment analysis. The first piece of work is a study I did with Jack Grieve, where we discovered that evaluation in reviews is mostly in the last third of the review (Taboada & Grieve 2004). By assigning a higher weight to words in the last third of the text, we improved the performance of our sentiment analysis system. We experimented with different weights, but we found that there was a peak of opinion-related adjectives at the two third mark of most reviews, which were short reviews posted online (Taboada 2008). The final weights were as shown in Figure 1. For instance, if the adjective *excellent* appeared in the first third of the text, its prior polarity (let us say, a positive 4) was multiplied by 0.2. This is because we had found that adjectives at the beginning of the text tended to constitute ‘noise’, i.e., they often referred to aspects other than the thing being reviewed. If the review was of a movie, the adjectives at the beginning may describe the mood of the reviewer at the time they went to see the movie or other contextual factors extraneous to the film itself. The word *excellent* could well refer to the concession popcorn. The beginning stages of a review tend to be descriptive rather than evaluative (Taboada et al. 2009). If the word *excellent*, however, appears towards the two third point in the text, then we multiply its weight by 1, assigning it the full dictionary value. This is a simple example of how application-oriented work led us to a linguistic insight about evaluative text, about the structure of stages in review texts and the distribution of descriptive vs. evaluative stages.

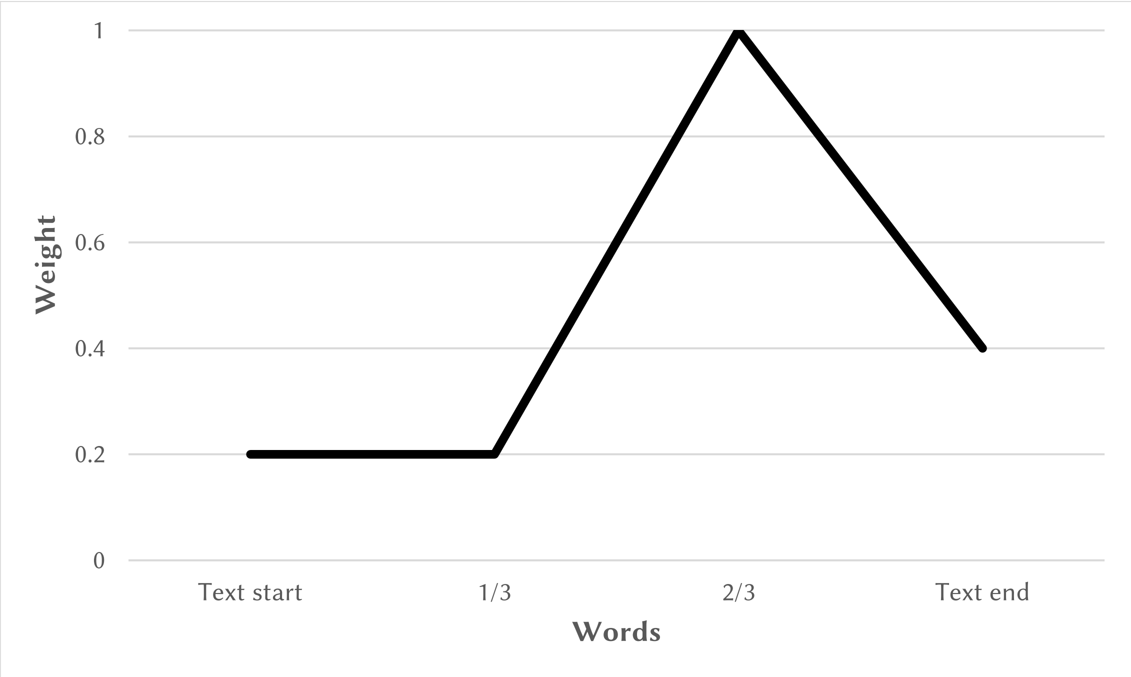

Figure 1: Weight distribution for evaluative adjectives in review texts

The second example of patterns of sentiment comes from other work examining the distribution of positive and negative evaluation. In Taboada & Gómez-González (2012), we found that, when we got things wrong by simply counting negative and positive words, it was because the positive was almost meant to be discounted. This we labelled as a vernacular style of argumentation, a “yes, but” kind of style, which Susan Hunston has also called a dialogue with no alternative viewpoints: “some people say X, but I say Y” (Hunston 2011). We found this pattern in concessive relations, which occurred either within the sentence, as in (8), or across sentences, structuring whole chunks of the discourse, as we see in (9). In the first example, two adjectives, *happy* and *sad* are present. If we were to average their values, we would conclude that this sentence conveys neutral evaluation. The author, however, is evaluating the ending as primarily sad, because the adjectives *sad* is found in the nucleus, the main part of this sentence.

(8) *Although the ending was a happy one, it was also a little sad.*

(9) *The good stuff....the visual production itself with its ultra-stylized appearance. It looks nice, but did the budget of a third world nation need to be spent to create this film?*

Finally, the third example of patterns of evaluative language is a pattern we discovered not at the discourse level, but at the phrase level, in the use of what we called *adverbly adjectives*, an adjective phrase made up of a deadjectival adverb (the *-ly* adverb) and an adjective, such as *deeply disturbing* or *delightfully performed*. We found that these are extremely common in evaluative text. They are quite interesting, because some of them are oxymorons, where one part is negative and the other positive. In a series of papers with Cliff Goddard and Radoslava Trnavac, we explored how this construction conveys evaluation (Goddard et al. 2016, 2018, 2019; Taboada et al. 2017; Trnavac & Taboada 2020; Taboada et al. 2024). We classified different combinations into semantic classes, conveying Degree, Focus, Manner, Reaction, Topical, and Epistemic. The most interesting, from our point of view, were the Reaction type, as they intensely conveyed evaluation, including in oxymorons like the ones shown in (10). They seem to be deployed as examples of linguistic virtuosity. Additionally, the construction is perhaps quite English specific: While it seems possible in other languages, it is not employed so frequently. What made this pattern challenging for sentiment analysis is that it is difficult to assign a numerical, compositional positive or negative label to the oxymorons. In an example like *devastatingly effective*, *devastating* is negative whereas *effective* is positive. It is difficult to decide what the value of the phrase should be.

(10) Reaction type: *incredibly good*, *breathtakingly simple*, *ridiculously complex*, *gorgeously written*, *comically predictable*

Oxymorons: *staggeringly incompetent*, *devastatingly effective*, *strangely compelling*, *beautifully heartbreaking*, *zealously moderate*

These three examples, the structure of reviews, vernacular argumentation, and *adverbly adjectives*, are all examples of phenomena we found in the process of doing NLP, that is, in the process of doing sentiment analysis. But they are all linguistically interesting and can perhaps be examples of what it means to do computational evaluation.

Many questions remain, about what we gain and lose with large-scale vs. qualitative analyses, and how to undertake each from a principled point of view. It is also quite likely that the boundaries, between large-scale and quantitative, between NLP and CL, do not matter as much, and the most important principle is that we are studying evaluative language as a phenomenon unto itself.

The last main section of this paper addresses another example of computational evaluation, that of the use of figurative language to convey evaluative meaning, and especially extremely negative evaluation of the type that appears in abusive and toxic language online.

4 Metaphors in abusive language online

In the context of finding abusive and hateful language online and doing so computationally, we find that, indeed, explicit abuse abounds online. Research has also shown, however, that abusive language sometimes wraps negative opinion in positive words (Taboada et al. 2017). I argue that metaphor is one of the ways we do this, because the pragmatic role of metaphor is usually to provide evaluation (Charteris-Black 2004). Some of the metaphors are quite conventionalized, in the form of idioms or appeals to common tropes, yet some of them are sophisticated and subtle uses of figurative language with the intention to offend.

Let us examine two examples of simple metaphor, without apparent abusive intent. The first is an example of a comment in response to a policy proposal by the NDP, one of Canada's political parties, to start a national daycare plan. The comment in (11) is about "growing money trees". This is somewhat conventionalized, an appeal to the idiom *money doesn't grow on trees*. Although not particularly offensive, the irony does intend to dismiss the contents of the article it refers to and the idea that the plan is financially feasible.[2] Other metaphors are somewhat more sophisticated; for instance, (12) refers to an increase in the number of elderly baby boomers as a tsunami. This appeals to the inevitable and destructive character of natural disasters (Dancygier 2016).

(11) *Wow, great idea! All we need to do now is to start* ***planting those money trees*** *and everything will be okay.*

(12) *the opposition parties fail to address the coming* ***economic Tsunami of aging boomers***

I would like to start by defining what I mean by abusive language. I will gloss over many important differences, but we can delimit slightly different phenomena that we talk about when we talk about abusive language. Some of it is offensive, abusive, or toxic, and some of it is hate speech. Hate speech is a category unto itself, mostly because it is well defined from a legal point of view, although different jurisdictions have different legislation on this, and what is allowed or prosecuted can vary (Leader Maynard & Benesch 2016; Brown 2017; Gill 2020). The other main category, apart from hate speech, includes terms like offensive, abusive, or toxic language (Waseem et al. 2017; Fortuna et al. 2020; Kennedy et al. 2022). In general, I prefer both *abusive*, which tends to evoke attack on a person, and *toxic*, which also conveys that toxic substances have a nasty habit of spreading. One reason platforms are encouraged to moderate toxic content

[2] https://www.theglobeandmail.com/opinion/daycare-picks-up-the-ndp/article21094039/, 12 July 2024

is because it poisons the space. The general consensus around toxic language is that there is no consensus (Avalle et al. 2024). Attempts at definitions, however, do have some common characteristics, having to do with the content (obscene, derogatory, violent, or extreme) and with the consequences (that the content is likely to make someone leave the conversation; Demjén & Hardaker 2016; Waseem et al. 2017; Sap et al. 2019; Kolhatkar et al. 2020; Vidgen & Derczynski 2021).

The challenge in automatic detection of abusive and toxic language is that metaphors are often deployed. Mendelsohn et al. (2020) examine the linguistic resources for dehumanizing and abusive language and conclude that they include metaphors related to animals, metaphors of disgust, and denial of subjectivity and agency. We could say, following Lakoff & Johnson (1980), that we not only live by metaphors but also hate by metaphors.

Let us examine some examples that illustrate this use of metaphors, but also other figures of speech, to convey abuse. The examples that follow come from a corpus of online news comments (Kolhatkar et al. 2020) and from a corpus of tweets directed at politicians in the 2019 Canadian federal election (Dubois & Owen 2019). In (13) and (14), we see examples of dehumanization. It is almost conventionalized to refer to people we do not like as animals, as we see in (13), where a politician is portrayed as a piglet. In the second example, the members of ISIS are labeled as Neanderthals, Cro-Magnons, or savages, all clearly meant derogatorily. In the examples, I have written in boldface the linguistic expressions that most overtly convey the metaphors.

(13) *The G & M refers to Mr. Kenney as a person of stature. In our house we refer to him as* ***piglet****. Often* ***up on his hind legs, squealing****.*

(14) *They say a guerrilla war is unwinnable. Bomb* ***these ISIS neanderthals into the stone age*** *and youll get a* ***hardier species of cro-magnon savages*** *that will make ISIS look like surreal tea party fairies. But, hey, enjoy. Give war a chance. Its good for bizness. And its really really good for show biz.*

The next set of examples illustrate a use of the metaphor 'bad person as a toxic substance'. (15) and (16) refer to Stephen Harper, who was, at the time of these comments in 2015, running as an incumbent to remain prime minister, an election that he eventually lost. A hashtag developed, *heave Steve*, which alluded to a desire to 'eject' him or 'get rid of him', which appealed to notions of toxicity and undesired stomach contents. (17) refers to another politician, this time very explicitly, as a *toxic idiot* and a source of disgust, another common metaphor for something undesirable.

(15) *Anybody But Harper. This man is a dangerous Power of One and we let him take it. Vote strategically to* ***heave Steve****. I refer the Liberal Program which includes returning to a more open government.*

(16) *CONS have no moral compass. Long past time to* ***#HeaveSteve****.*

(17) *@MichelleRempel @CreativeTweets Ur a* ***toxic idiot*** *Michelle and totally irresponsible. U* ***disgust*** *me*

Beyond metaphors, other rhetorical devices and figures of speech are used to convey negative, abusive, or toxic opinions. By rhetorical devices I mean the traditional, classical sense of figures and tropes. For instance, rhetorical questions and reported speech play a role in abusive language. Rhetorical questions, questions without an answer or where the answer is obvious, are often deployed. Quoting somebody is often a way of evoking what they said, and it can be used to distance the speaker from the content or to mock it. (18) below illustrates a repeated use of rhetorical questions, whereas (19) uses direct speech (*What is wrong with this site???*), i.e., speech by another poster, to mock that poster.

(18) *Less is done? can you actually support that claim with real evidence? Or have you just swallowed the hype, hook, line and sinker?*

(19) *"What is wrong with this site???" - that it lets users like you register and ask questions.*

Other figures of speech, such as euphemism, litotes, sarcasm, or antithesis are often deployed. One of the most productive resources for abusive language is sarcasm and irony. This brings us back to the issue that sarcasm and irony are a very difficult problem for sentiment analysis. They will also be a difficult problem in abusive language detection. It is also important to remember is that these rhetorical devices often come together, and it is the combination that produces a pile-on effect and the sense that the message is abusive (Charteris-Black 2005).

We can treat this as a computational evaluation problem. We may be able to find metaphors using NLP, but what I am interested in is whether we can identify the ones used for abuse and whether this line of research will reveal something interesting about abusive language itself. Research on automatic detection of metaphors is thriving and there is hope that accurate detection of various types is not far off (Shutova et al. 2013; Harris et al. 2018; Tong et al. 2021; Kühn & Mitrović 2024). Other subfields of linguistics, including corpus linguistics, corpus-based discourse analysis, and *appraisal*, are also contributing to the problem of identifying metaphors, computationally or not (Charteris-Black 2004; Partington et al. 2004; Martin & White 2005; Semino 2008; Dancygier & Sweetser 2014).

5 Conclusion

I propose in this paper that we can view computational evaluation as a form of linguistics and of CL, a field that has as a goal to represent and understand evaluative language. I should also acknowledge that I am not defining a whole new field, but rather labelling and trying to bring together under the computational evaluation label the type of research that many scholars are already engaged in.

The other main conclusion I draw is that, indeed, metaphors and other figures of speech play a powerful role in conveying abuse and hate and that we are still working on how to detect them automatically, but I believe that, along that way, we will also derive interesting insights about language itself.

Acknowledgements: This research is funded by the Natural Sciences and Engineering Research Council of Canada. I would like to thank the audience at the Conference of the German Society for Linguistics (DGfS) in Bochum in February 2024 for excellent questions and feedback.

References

Asr, Fatemeh Torabi, Mehrdad Mokhtari & Maite Taboada. 2023. Misinformation detection in news text: Automatic methods and data limitations. In Stefania M. Maci, Massimiliano Demata, Mark McGlashan & Philip Seargeant (eds.), *The Routledge handbook of discourse and disinformation*, 79–102. New York: Routledge. https://doi.org/10.4324/9781003224495-7.

Asr, Fatemeh Torabi & Maite Taboada. 2019. Big data and quality data for fake news and misinformation detection. *Big Data and Society* 6. 1–14. https://doi.org/10.1177/2053951719843310.

Avalle, Michele, Niccolò Di Marco, Gabriele Etta, Emanuele Sangiorgio, Shayan Alipour, Anita Bonetti, Lorenzo Alvisi, Antonio Scala, Andrea Baronchelli & Matteo Cinelli. 2024. Persistent interaction patterns across social media platforms and over time. *Nature* 628. 582–589. https://doi.org/10.1038/s41586-024-07229-y.

Babu, Nirmal Varghese & E. Grace Mary Kanaga. 2022. Sentiment analysis in social media data for depression detection using artificial intelligence: A review. *SN Computer Science* 3. 1–20. https://doi.org/10.1007/s42979-021-00958-1.

Badia, Toni, Lingxian Bao & Patrik Lambert. 2024. Improving aspect-based neural sentiment classification with lexicon enhancement, attention regularization and sentiment induction. *Natural Language Engineering* 30(1). 1–30. https://doi.org/10.1017/S1351324922000432.

Balahur, Alexandra, Ralf Steinberger, Mijail Kabadjov, Vanni Zavarella, Erik Van Der Goot, Matina Halkia, Bruno Pouliquen & Jenya Belyaeva. 2010. Sentiment analysis in the news. *Language Resources and Evaluation Conference* (LREC) 7. 2216–2220.

Banfield, Ann. 1982. *Unspeakable Sentences: Narration and representation in the language of fiction*. Boston: Routledge and Kegan Paul.

Bansal, Mohit, Claire Cardie & Lillian Lee. 2008. The power of negative thinking: Exploiting label disagreement in the min-cut classification framework. *Posters and Demonstrations at the International Conference on Computational Linguistics* (Coling) 22. 15–18.

Bednarek, Monika. 2008. *Emotion talk across corpora*. New York: Palgrave Macmillan. https://doi.org/10.1057/9780230285712.

Benamara, Farah, Diana Inkpen & Maite Taboada. 2018. Introduction to the special issue on language in social media: Exploiting discourse and other contextual information. *Computational Linguistics* 44(4). 663–681. https://doi.org/10.1162/coli_a_00333.

Benamara, Farah, Maite Taboada & Yannick Mathieu. 2017. Evaluative language beyond bags of words: Linguistic insights and computational applications. *Computational Linguistics* 43(1). 201–264. https://doi.org/10.1162/COLI_a_00278.

Berwick, Robert C. & Edward P. Stabler. 2019. *Minimalist parsing*. Oxford: Oxford University Press. https://doi.org/10.1093/oso/9780198795087.001.0001.

Biber, Douglas & Edward Finegan. 1988. Adverbial stance types in English. *Discourse Processes* 11(1). 1–34. https://doi.org/10.1080/01638538809544689.

Blackburn, Patrick & Johan Bos. 2003. Computational semantics. *Theoria: An International Journal for Theory, History and Foundations of Science* 18. 27–45. https://doi.org/10.1387/theoria.408.

Bouktif, Salah, Ali Fiaz & Mamoun Awad. 2020. Augmented textual features-based stock market prediction. *IEEE Access* 8. 40269–40282. https://doi.org/10.1109/ACCESS.2020.2976725.

Brown, Alexander. 2017. What is hate speech? Part 1: The myth of hate. *Law and Philosophy* 36(4). 419–468. https://doi.org/10.1007/s10982-017-9297-1.

Canute, Matt, Mali Jin, Alberto Lusoli, Philippa R Adams, Mugdha Pandya, Maite Taboada, Diana Maynard & Wendy Hui Kyong Chun. 2023. Dimensions of online conflict: Towards modeling agonism. *Findings of the Association for Computational Linguistics: Empirical Methods in Natural Language Processing* (EMNLP) 2023. 12194–12209. https://doi.org/10.18653/v1/2023.findings-emnlp.816.

Cavasso, Luca & Maite Taboada. 2021. A corpus analysis of online news comments using the appraisal framework. *Journal of Corpora and Discourse Studies* 4(1). 1–38. https://doi.org/10.18573/jcads.61.

Chafe, Wallace & Johanna Nichols. 1986. *Evidentiality: The linguistic coding of epistemology*. Norwood: Ablex.

Charteris-Black, Jonathan. 2004. *Corpus approaches to critical metaphor analysis*. New York: Palgrave Macmillan. https://doi.org/10.1057/9780230000612.

Charteris-Black, Jonathan. 2005. *Politicians and rhetoric: The persuasive power of metaphor*. New York: Palgrave Macmillan. https://doi.org/10.1057/9780230501706.

Cui, Jing, Zhaoxia Wang, Seng-Beng Ho & E. Cambria. 2023. Survey on sentiment analysis: Evolution of research methods and topics. *Artificial Intelligence Review* 56. 8469–8510. https://doi.org/10.1007/s10462-022-10386-z.

Dancygier, Barbara. 2016. Figurativeness, conceptual metaphor, and blending. In Elena Semino & Zsófia Demjén (eds.), *The Routledge handbook of metaphor and language*, 28–41. New York: Routledge.

Dancygier, Barbara & Eve Sweetser. 2012. *Viewpoint in language: A multimodal perspective*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139084727.

Dancygier, Barbara & Eve Sweetser. 2014. *Figurative language*. Cambridge: Cambridge University Press.

De Marneffe, Marie-Catherine, Christopher D. Manning, Joakim Nivre & Daniel Zeman. 2021. Universal dependencies. *Computational Linguistics* 47(2). 255–308. https://doi.org/10.1162/coli_a_00402.

Demjén, Zsófia & Claire Hardaker. 2016. Metaphor, impoliteness, and offence in online communication. In Elena Semino & Zsófia Demjén (eds.), *The Routledge handbook of metaphor and language*, 353–367. New York: Routledge.

Donahue, Evan, Matt Canute, Adjua Akinwumi, Philippa R. Adams, Maite Taboada & Wendy Hui Kyong Chun. 2025. Algorithm obscurity in bossware: A sentiment analysis case study. Manuscript in review at *AI & Society*.

Dubois, Elizabeth & Taylor Owen. 2019. *Understanding the digital ecosystem: Findings from the 2019 federal election*. Ottawa: McGill University and University of Ottawa report.

Ein-Dor, Liat, Ilya Shnayderman, Artem Spector, Lena Dankin, Ranit Aharonov & Noam Slonim. 2022. Fortunately, discourse markers can enhance language models for sentiment analysis. *Conference on Artificial Intelligence by the Association for the Advancement of Artificial Intelligence* (AAAI) 36. 10608–10617. https://doi.org/10.1609/aaai.v36i10.21305.

Fortuna, Paula, Juan Soler & Leo Wanner. 2020. Toxic, hateful, offensive or abusive? What are we really classifying? An empirical analysis of hate speech datasets. *Language Resources and Evaluation Conference* (LREC) 12. 6786–6794.

Fruehwald, Josef. 2024. *What is NLP?* https://jofrhwld.github.io/teaching/courses/2022_lin517/lectures/what_is_nlp/, 25 February 2025.

Giannakidou, Anastasia. 1995. On the semantic licensing of polarity items. *Studies in Greek Linguistics* 15. 406–418.

Giannakidou, Anastasia. 2001. Varieties of polarity items and the (non)veridicality hypothesis. In Jack Hoeksema, Hotze Rullmann, Víctor Sánchez-Valencia & Ton van der Wouden (eds.), *Perspectives on negation and polarity items*, 99–127. Amsterdam: John Benjamins. https://doi.org/10.1075/la.40.06gia.

Gill, Lex. 2020. *The legal aspects of hate speech in Canada*. Ottawa: Public Policy Forum report.

Gillespie, Tarleton. 2020. Content moderation, AI, and the question of scale. *Big Data and Society* 7(2). 1–5. https://doi.org/10.1177/2053951720943234.

Global Markets Training. 2024. *ChatGPT for equity sentiment analysis*. https://www.globalmarkets-training.co.uk/training-courses/using-llms-for-sentiment-analysis/, 16 May 2024.

Goddard, Cliff, Maite Taboada & Radoslava Trnavac. 2016. *Semantic descriptions of 24 evaluational adjectives, for application in sentiment analysis*. https://arxiv.org/abs/1608.06697, 25 February 2025.

Goddard, Cliff, Maite Taboada & Radoslava Trnavac. 2018. Hilariously ridiculous and other adverb-adjective combinations: Classification and frequency distribution across registers. Paper presented at the *American Pragmatics Association* (AMPRA) 4, University at Albany, 1–3 November.

Goddard, Cliff, Maite Taboada & Radoslava Trnavac. 2019. The semantics of evaluational adjectives: Perspectives from natural semantic metalanguage and Appraisal. *Functions of Language* 26(3). 308–342. https://doi.org/10.1075/fol.00029.god.

Guo, Zhijiang, Michael Schlichtkrull & Andreas Vlachos. 2022. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics* (TACL) 10. 178–206. https://doi.org/10.1162/tacl_a_00454.

Gutzmann, Daniel. 2019. *The grammar of expressivity*. Oxford: Oxford University Press. https://doi.org/10.1093/oso/9780198812128.001.0001.

Harris, Randy Allen, Chrysanne Di Marco, Sebastian Ruan & Cliff O'Reilly. 2018. An annotation scheme for rhetorical figures. *Argument & Computation* 9(2). 155–175. https://doi.org/10.3233/AAC-180037.

Hoang, Mickel, Oskar Alija Bihorac & Jacobo Rouces. 2019. Aspect-based sentiment analysis using BERT. *Nordic Conference on Computational Linguistics* (NoDaLiDa) 22. 187–196.

Huber, Patrick & Giuseppe Carenini. 2022. Towards understanding large-scale discourse structures in pre-trained and fine-tuned language models. *North American Chapter of the Association for Computational Linguistics* (NAACL) 2022. 2376–2394. https://doi.org/10.18653/v1/2022.naacl-main.170.

Hunston, Susan. 2011. *Corpus approaches to evaluation: Phraseology and evaluative language*. New York: Routledge. https://doi.org/10.4324/9780203841686.

Jurafsky, Daniel & James H. Martin. 2024. *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. 3rd edn. Upper Saddle River: Prentice Hall.

Kennedy, Brendan, Mohammad Atari, Aida Mostafazadeh Davani, Leigh Yeh, Ali Omrani, Yehsong Kim, Kris Coombs, Shreya Havaldar, Gwenyth Portillo-Wightman & Elaine Gonzalez. 2022. Introducing the Gab Hate Corpus: Defining and applying hate-based rhetoric to social media posts at scale. *Language Resources and Evaluation* 56. 79–108. https://doi.org/10.1007/s10579-021-09569-x.

Kim, Soo-Min & Eduard Hovy. 2006. Automatic identification of pro and con reasons in online reviews. *Main Conference Poster Session at the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics* (Coling/ACL) 2006. 183–490. https://doi.org/10.3115/1273073.1273136.

Kim, Soo-Min & Eduard Hovy. 2007. Crystal: Analyzing predictive opinions on the web. *Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (EMNLP-CoNLL) 2007. 1056–1064.

Kolhatkar, Varada, Nithum Thain, Jeffrey Sorensen, Lucas Dixon & Maite Taboada. 2023. Classifying constructive comments. *First Monday* 28(4). https://doi.org/10.5210/fm.v5228i5214.13163.

Kolhatkar, Varada, Hanhan Wu, Luca Cavasso, Emilie Francis, Kavan Shukla & Maite Taboada. 2020. The SFU Opinion and Comments Corpus: A corpus for the analysis of online news comments. *Corpus Pragmatics* 4(2). 155–190. https://doi.org/10.1007/s41701-019-00065-w.

Ku, Lun-Wei, Yu-Ting Liang & Hsin-Hsi Chen. 2006. Opinion extraction, summarization and tracking. *Spring Symposium on Computational Approaches to Analyzing Weblogs by the Association for the Advancement of Artificial Intelligence* (AAAI) 2006. 100–107.

Kühn, Ramona & Jelena Mitrović. 2024. The elephant in the room: Ten challenges of computational detection of rhetorical figures. *Figurative Language Processing* (FigLang) 4. 45–52. https://doi.org/10.18653/v1/2024.figlang-1.6.

Lakoff, George & Mark Johnson. 1980. *Metaphors we live by*. Chicago: University of Chicago Press.

Leader Maynard, Jonathan & Susan Benesch. 2016. Dangerous speech and dangerous ideology: An integrated model for monitoring and prevention. *Genocide Studies and Prevention: An International Journal* 9(3). 70–95. https://doi.org/10.5038/1911-9933.9.3.1317.

Loáiciga, Sharid, Anne Beyer & David Schlangen. 2022. New or old? Exploring how pre-trained language models represent discourse entities. *International Conference on Computational Linguistics* (Coling) 29. 875–886.

Malouf, Robert & Tony Mullen. 2008. Taking sides: Graph-based user classification for informal online political discourse. *Internet Research* 18. 177–190. https://doi.org/10.1108/10662240810862239.

Martin, James R. & Peter R. R. White. 2005. *The language of evaluation*. New York: Palgrave.

Mendelsohn, Julia, Yulia Tsvetkov & Dan Jurafsky. 2020. A framework for the computational linguistic analysis of dehumanization. *Frontiers in Artificial Intelligence* 3(55). https://doi.org/10.3389/frai.2020.00055.

Mohammad, Saif M., Xiaodan Zhu, Svetlana Kiritchenko & Joel Martin. 2015. Sentiment, emotion, purpose, and style in electoral tweets. *Information Processing and Management* 51(4). 480–499. https://doi.org/10.1016/j.ipm.2014.09.003.

Niu, Yun, Xiaodan Zhu, Jianhua Li & Graeme Hirst. 2005. Analysis of polarity information in medical text. *American Medical Informatics Association* (AMIA) 2005. 570–574.

Pang, Bo, Lillian Lee & Shivakumar Vaithyanathan. 2002. Thumbs up? Sentiment classification using machine learning techniques. *Empirical Methods in Natural Language Processing* (EMNLP) 7. 79–86. https://doi.org/10.3115/1118693.1118704.

Partington, Alan, John Morley & Louann Haarman (eds.). 2004. *Corpora and discourse*. Bern: Peter Lang.

Raskin, Victor. 1987. Linguistics and natural language processing. *Theoretical and Methodological Issues in Machine Translation of Natural Languages* (TMIMTNL) 1. 42–58.

Roccabruna, Gabriel, Steve Azzolin & Giuseppe Riccardi. 2022. Multi-source multi-domain sentiment analysis with BERT-based models. *Language Resources and Evaluation Conference* (LREC) 13. 581–589.

Sap, Maarten, Dallas Card, Saadia Gabriel, Yejin Choi & Noah A. Smith. 2019. The risk of racial bias in hate speech detection. *Association for Computational Linguistics* (ACL) 57. 1668–1678. https://doi.org/10.18653/v1/P19-1163.

Semino, Elena. 2008. *Metaphor in discourse*. Cambridge: Cambridge University Press.

Shutova, Ekaterina, Simone Teufel & Anna Korhonen. 2013. Statistical metaphor processing. *Computational Linguistics* 39(2). 301–353. https://doi.org/10.1162/COLI_a_00124.

Somasundaran, Swapna & Janyce Wiebe. 2009. Recognizing stances in online debates. *Association for Computational Linguistics* (ACL) 47. 226–234. https://doi.org/10.3115/1687878.1687912.

Stab, Christian & Iryna Gurevych. 2017. Parsing argumentation structures in persuasive essays. *Computational Linguistics* 43(3). 619–659. https://doi.org/10.1162/COLI_a_00295.

Stanford, James N. & Laurence A. Kenny. 2013. Revisiting transmission and diffusion: An agent-based model of vowel chain shifts across large communities. *Language Variation and Change* 25(2). 119–153. https://doi.org/10.1017/S0954394513000069.

Taboada, Maite. 2008. SFU Review Corpus. Vancouver: Simon Fraser University corpus.

Taboada, Maite. 2016. Sentiment analysis: An overview from linguistics. *Annual Review of Linguistics* 2. 325–347. https://doi.org/10.1146/annurev-linguistics-011415-040518.

Taboada, Maite, Julian Brooke & Manfred Stede. 2009. Genre-based paragraph classification for sentiment analysis. *Special Interest Group on Discourse and Dialogue* (SIGDIAL) 10. 62–70. https://doi.org/10.3115/1708376.1708385.

Taboada, Maite, Julian Brooke, Milan Tofiloski, Kimberly Voll & Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational Linguistics* 37(2). 267–307. https://doi.org/10.1162/COLI_a_00049.

Taboada, Maite & María de los Ángeles Gómez-González. 2012. Discourse markers and coherence relations: Comparison across markers, languages and modalities. *Linguistics and the Human Sciences* 6(1–3). 17–41. https://doi.org/10.1558/lhs.v6i1-3.17.

Taboada, Maite, Cliff Goddard & Radoslava Trnavac. 2024. The 'adverb-ly adjective' construction in English: Meanings, distribution, and discourse functions. *English Language & Linguistics* (online). 1–30. https://doi.org/10.1017/S136067432400008X.

Taboada, Maite & Jack Grieve. 2004. Analyzing appraisal automatically. *Spring Symposium on Exploring Attitude and Affect in Text by the Association for the Advancement of Artificial Intelligence* (AAAI) 2004. 158–161.

Taboada, Maite, Radoslava Trnavac & Cliff Goddard. 2017. On being negative. *Corpus Pragmatics* 1(1). 57–76. https://doi.org/10.1007/s41701-017-0006-y.

Tong, Xiaoyu, Ekaterina Shutova & Martha Lewis. 2021. Recent advances in neural metaphor processing: A linguistic, cognitive and social perspective. *North American Chapter of the Association for Computational Linguistics* (NAACL) 2021. 4673–4686. https://doi.org/10.18653/v1/2021.naacl-main.372.

Trnavac, Radoslava & Maite Taboada. 2020. Positive appraisal in online news comments. In Kerry Mullan, Bert Peeters & Lauren Sadow (eds.), *Studies in ethnopragmatics, cultural semantics, and intercultural communication: Ethnopragmatics and semantic analysis*, 185–205. Singapore: Springer Singapore. https://doi.org/10.1007/978-981-32-9983-2_10.

Tsujii, Junichi. 2021. Natural language processing and computational linguistics. *Computational Linguistics* 47(4). 707–727. https://doi.org/10.1162/coli_a_00420.

Vidgen, Bertie & Leon Derczynski. 2021. Directions in abusive language training data, a systematic review: Garbage in, garbage out. *PLoS One* 15(12). 1–32. https://doi.org/10.1371/journal.pone.0243300.

Waseem, Zeerak, Thomas Davidson, Dana Warmsley & Ingmar Weber. 2017. Understanding abuse: A typology of abusive language detection subtasks. *Workshop on Abusive Language Online* (ALW) 1. 78–84. https://doi.org/10.18653/v1/W17-3012.

Wiebe, Janyce. 1994. Tracking point of view in narrative. *Computational Linguistics* 20(2). 233–287.

Wiebe, Janyce, Theresa Wilson, Rebecca Bruce, Matthew Bell & Melanie Martin. 2004. Learning subjective language. *Computational Linguistics* 30(3). 277–308. https://doi.org/10.1162/0891201041850885.

Wierzbicka, Anna. 1999. *Emotions across languages and cultures*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511521256.

Wilson, Theresa, Janyce Wiebe & Paul Hoffmann. 2009. Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis. *Computational Linguistics* 35(3). 399–433. https://doi.org/10.1162/coli.08-012-R1-06-90.

Xie, Jie, Ri Le Ge Su & Jaehoon Song. 2022. An analytical study of employee loyalty and corporate culture satisfaction assessment based on sentiment analysis. *Frontiers in Psychology* 13. 1–9. https://doi.org/10.3389/fpsyg.2022.971569.

Zucco, Chiara, Barbara Calabrese & Mario Cannataro. 2017. Sentiment analysis and affective computing for depression monitoring. *IEEE International Conference on Bioinformatics and Biomedicine* (BIBM) 2017. 1988–1995. https://doi.org/10.1109/BIBM.2017.8217966.