# The Linguistic Ideologies of Deep Abusive Language Classification

**Michael Castelle**
Centre for Interdisciplinary Methodologies
University of Warwick
M.Castelle.1@warwick.ac.uk

## Abstract

This paper brings together theories from sociolinguistics and linguistic anthropology to critically evaluate the so-called "language ideologies"—the set of beliefs and ways of speaking about language—in the practices of abusive language classification in modern machine learning-based NLP. This argument is made at both a conceptual and empirical level, as we review approaches to abusive language from different fields, and use two neural network methods to analyze three datasets developed for abusive language classification tasks (drawn from Wikipedia, Facebook, and StackOverflow). By evaluating and comparing these results, we argue for the importance of incorporating theories of pragmatics and metapragmatics into both the design of classification tasks as well as in ML architectures.

## 1 Introduction

> Some problems lend themselves more easily to A.I. solutions than others. So, hate speech is one of the hardest, because determining if something is hate speech is very linguistically nuanced. Right?... I'm optimistic that over a five- to 10-year period, we will have A.I. tools that can get into some of the nuances, the linguistic nuances of different types of content to be more accurate in flagging things for our systems.
>
> (Excerpt from testimony to the U.S. Senate and Judiciary and Commerce, Science and Transportation Committees by Mark Zuckerberg, April 10th, 2018)

The term *nuance*, as used in the above quote from Mark Zuckerberg's testimony to the U.S. Senate, appears to imply that, like the successes of convolutional neural networks in computer vision, the classification (or "flagging") of (textual) hate speech should be considered a matter of advanced pattern recognition, of recognizing the right 'shade' or 'color' of a given sentence in isolation. This view of linguistic action is in contrast with a tradition of speech-act theory (Austin, 1962; Searle, 1969) in which the meaning of an utterance is not solely determined by lexical or syntactic structure but by its *social context* and *performative effects*. In studies of hate speech following these latter perspectives, such as those of Butler (1997), words are seen not just as arbitrary lexemes but something capable of actively causing violence to one or more addressees.

Butler, in her account of hate speech, optimistically saw a 'gap' between a speech act and its effects (Butler, 1997)—comparable to Austin's distinction between the *illocutionary* (the addressee's intention) and *perlocutionary* (the outcome of the utterance)—which she uses to argue against legal regulation of hate speech and instead for the possibility of 'restaging' and 'resignifying' such speech on the part of the addressee. Schwartzman (2002), however, contended that Butler's account was fundamentally missing an awareness of existing social structures of power, and that if any successful 'resignification' (as in the "taking back" of slurs) occurs it is due to an active and political response against oppression.

The question remains: is detecting hate speech a matter of sufficiently advanced pattern recognition a matter of building the right distributed architecture on a large enough data set of "hateful" vs. "non-hateful" expressions? Or is it necessary for any such automated classification to be integrated with long-term ethnographic and contextual immersion in the communities, however problematic, in which said expressions emerge (Sinders and Martinez, 2018)? To what extent should



*Proceedings of the Second Workshop on Abusive Language Online (ALW2)*, pages 160–170
Brussels, Belgium, October 31, 2018. ©2018 Association for Computational Linguistics

we expect researchers to attempt to integrate the sociocultural complexities of contemporary online communication into their "A.I. tools", and to what extent should we expect success or failure from those tools?

This article assesses the current potential for, and limitations of, machine learning methodologies—both in terms of how the 'gold standard' datasets are normally constructed and how neural networks are applied—for abusive language detection in natural language text. We will, through our own experiments, explore the specific conflicts between the research culture of achieving "state-of-the-art" scores (Manning, 2015) with (relatively) black-boxed architectures and theories of the pragmatics (and metapragmatics) of linguistic abuse in practice.

## 2 The Language Ideologies of NLP

With its 1950s intellectual origins in a) naïve, word-by-word machine translation and b) Chomsky's conception of language as generated by a machinic automata, the field of natural language processing (NLP) inherited what we refer to as a *language ideology* (Woolard and Schieffelin, 1994), a set of beliefs and ways of speaking about 'language'.[1] Fundamental to the Chomskyan language ideology, for example, is the purported existence of an innate language faculty which is held to explain the purported lack of training data for humans (the so-called "poverty of the stimulus") (Pereira, 2000). This faculty is conceived as the locus of a *generative grammar* which can in theory enumerate all possible valid sentences.

Other disciplines of human communication, however, developed alternative perspectives which embed language in a more complex sociocultural environment. The field of *sociolinguistics* (Labov, 1972) considered quantitative measurements of language variation (such as dialects and creoles) not as a matter of syntactic validity but as intrinsically related to social differentiation and class structure; and *linguistic anthropology* (Duranti, 1997) simultaneously drew from the semiotic and structuralist traditions of Peirce and Jakobson, emphasizing acts of reflexive, sociocultural entextualization and contextualization (Bauman and Briggs, 1990) inherent to any communicative situation.

From the perspective of these latter fields, the study of 'natural language' by computer scientists (so named to distinguish itself from formal logic and 'artificial' programming languages devised in early AI research) would be seen as largely focused on *entextualized* utterances (such as written sentences or recorded speech) and the uncovering (or 'processing') of their hidden symbolic patterns and structures. The introduction of *statistical* approaches to NLP (Charniak, 1996) laid the groundwork for the eventual incorporation of a machine learning methodology which uses of a *corpus* of data composed of training inputs with labeled outputs. Today's neural network models in NLP are strongly dependent on this arguably *behaviorist* (i.e. stimulus-response) ideology of language, one in which syntax and semantics can be durably encoded through the presentation and re-presentation of vectorized 'stimuli' (which, through backpropagation, modifies the model's parameters based on the distance of the 'response' vector from the true values).

The study of abusive language in NLP, then, represents a profitable collision between two potentially compatible language ideologies — one a "statistical NLP" ideology which claims the potential to efficiently and intelligently discretize and classify the contextually dense and multimodal miasma of real-time communication; and the other, to be described in the next section, which seeks to carve out and eliminate impurities and danger (e.g., abusive language) in search of a 'safe', non-'toxic', and yet highly scalable environment.

## 3 Abusive Language Ideologies

So, if *language ideologies* are a set of beliefs and ways of speaking about 'language', then *abusive language ideologies* are a set of beliefs and ways of speaking about what it means for 'language' to be 'abusive'. As a review of the literature on this topic immediately indicates, there are a variety of theories regarding the nature of hate speech, abusive language, cyberbullying, etc.; in this section we will characterize the main positions, especially in their potential relation to NLP methodologies.

[1] The term 'ideology' as it used here is not intended as a pejorative; following Woolard (1998), we see ideology as "derived from, rooted in, reflective of, or responsive to the experience or interests of a particular social position" but not necessarily "in the service of the struggle to acquire or maintain power" (although it may also be that.)

### 3.1 Politeness

The study of online antagonism was preceded by much research in the linguistic field of pragmatics on *politeness* such as Brown and Levinson (1987), who isolated the concept of politeness as a set of interactional strategies to preserve 'face', a concept from Goffman (1967) reflecting the ideal self-image of the addresser or addressee in each communicative situation. This approach took into account the four cooperative 'maxims' of Grice (1975), which are implicit interactional norms in which speakers strive to be informative, unambiguous, brief, and orderly—norms which, we suggest, are potentially relevant to understanding online Q&A platforms like StackOverflow. Culpeper (1996) showed how a focus on *impoliteness* brought the importance of interactional context to the fore, in contrast to a focus on surface form in the work of Brown and Levinson; but later work on gender difference and politeness argued that impoliteness itself is only something classified as such by those in dominant positions of power (Mills, 2003). The linguistic study of politeness thus helpfully charts a development from the study of lexical and syntactic structure to interactional pragmatics to considerations of power relations within and among communities of practice.

### 3.2 Hate Speech

The concept of hate speech (and the debates surrounding its definition, legal and otherwise) is itself predicated on precisely a (conscious or unconscious) philosophical dispute about whether language can be segregated from action; the "fighting words" doctrine in U.S. law (Chaplinsky vs. New Hampshire, 1942), for example, was a legal intervention that (in a single instance) outlawed speech acts capable of provoking violent action (i.e. based on their perlocutionary force). From such an Austinian perspective, to be a free speech absolutist is to consider speech merely as locution and not illocution or perlocution (Hornsby and Langton, 1998); NLP methodology, which typically takes as input a set of text-artifacts segregated from their original communicative situation and consequences, could also be said to (implicitly) take this position of considering speech solely as locution (even if researchers commonly appreciate that their data was drawn from a past existence in richer contexts). For example, the work of Warner and Hirschberg (2012) acknowledges the limits of their decontextualized comment dataset, but argues that hate speech can still be distinguished through the recognition of stereotyped expressions about social groups.

The most comprehensive philosophical attempt to give meaning to the concept of hate speech is by Alexander Brown. In his two recent articles (Brown (2017a), Brown (2017b)), he explains that the term 'hate speech' likely emerged from a 1988 conference paper at Hofstra (Matsuda, 1989), and came to take on significant value for legal scholars before becoming integrated into a more popular discourse. Matsuda's intervention represents a *performative ideology* of language; if words are action, and some words are violent action, then hate speech can be regulated without violating a constitutional free speech principle. However, Brown comes to a somewhat negative conclusion regarding the possibility of coming to a coherent universal definition of the concept of hate speech; while it can be summarized as "a rough but nevertheless serviceable term to describe... the expressive dimensions of identity-based envy, hostility, conflict, mistrust and oppression", more fine-grained enclosures are never sufficient, and he argues that the meaning of 'hate speech' is closer to the 'family resemblances' concept of Wittgenstein (1953), who argues that terms like 'game' (and—implicitly—'language') can only denote an ever-shifting family of related practices irreducable to a precise definition.

It is thereby unsurprising that NLP's methodological detachment from the speech situation (itself embedded in other sociocultural contexts which do not become part of the training data), along with the fundamental indeterminacy of the 'hate speech' category, makes the reliability of 'coding' for hate speech a significant challenge (Ross et al., 2017). In their survey of hate speech detection in NLP, Schmidt and Wiegand (2017) mention this time-consuming dependency on hand-labeled data; but they also point out the many strategic ways that NLP researchers have proposed (if not always attempted) to overcome the decontextualized limitations and problems of definition of their data, and we will report similar findings below.

### 3.3 Abusive Language

Despite his critical conclusions, Brown argues that the term 'hate speech' is still, for now, effective; it "is used because it is useful, and it will remain useful so long as it can be used to do more than merely signal disapproval. If [that was] all it did... [it] would soon fall out of fashion or be replaced by newer, cooler bits of language that did the same thing but in more interesting ways" (Brown, 2017a). One such current variant term is *abusive language*, which appears in some of the earliest literature on antagonistic Internet communication (Spertus, 1997) but has in the past years taken on a greater prominence.

In part, the potentially milder connotations of the 'abusive language' term reflects a shifting from seeing online abuse as occurring on behalf of identity-based communities to occurring towards social groups in general, where those social groups might be something like "new users of StackOverflow". So for example, Nobata et al. (2016) uses the concept of abusive language to include hate speech, derogatory language, and profanity together. In their work on personal attacks in Wikipedia talk pages, Wulczyn et al. (2016) adopts the rhetoric of 'toxic' behavior, a term which metaphorically transposes affective concepts (such as *hate*) to one of environmental contamination and taboo (Douglas, 1966; Nagle, 2009); this represents a subtle move away from an otherwise dominant *personalist ideology* in which meaning emerges from the beliefs or intentions of the speaker (Hill, 2008).

Recognizing the overall lack of consensus on the boundaries of abusive language, Waseem et al. (2017) proposes a twofold typology: (1) whether language is "directed towards a specific individual or entity" or "directed towards a generalized group" and (2) whether the content is 'explicit' or 'implicit'. The resulting four axes, then, are each analyzed for the methodological approaches needed. *Directed* abuse can be detected with attention to proper nouns and entities like usernames; *Generalized* abuse may be associated with lexical patterns based on the targeted groups; *Explicit* abuse also often involves specific keywords and *Implicit* abuse the most difficult category, where more advanced semantic approaches such as word embeddings can fail in a complex polysemous and creative environment.

The view that indirectness and implicitness in text-artifacts can be eventually 'captured' by machine learning models is related to the performative ideology of speech-act theory, which has been criticized for its overemphasis on so-called *explicit* performatives (such as "I now pronounce you man and wife") over (far more common) implicit performative utterances which depend on contextual cues (Gumperz, 1982; Lempert, 2012). As Asif Agha puts it, "an indirect speech act is just a name for the way in which a denotational text diagrams an interactional text without describing it" (Agha, 2007, p. 100). This diagramming instead often happens through forms of pragmatic and metapragmatic *reflexivity* which may be difficult to recognize through analyzing the utterance detached from its interactional context, as is often the case for NLP datasets.

That researchers see indirect and implicit speech as a significant challenge, however, is in part due to our methodological embeddedness in a *referentialist ideology* which typically holds that the meaning of words are stable (as realized, for example, by static embedding vectors), and that the purpose of language is to communicate information. Jane Hill explains how the combination of referentialist and performative ideologies typifies conventional approaches to racism:

> Stereotypes and slurs are visible as racist to most people because they are made salient by referentialist and performative linguistic ideologies respectively. But other kinds of talk and text that are not visible, so called covert racist discourse, may be just as important in reproducing the culturally shared ideas that underpin racism. Indeed, they may be even more important, because they do their work while passing unnoticed. (Hill, 2008)

We argue that it will be essential for NLP researchers to recognize how our tools and techniques may, in part, be material embodiments of these ideologies, but also how one might partially escape those ideologies without abandoning the use of tools and techniques entirely. One positive example of this is from Saleem et al. (2017), which argues that supervised labeling based on keywords is problematic, but also that one can improve performance by training on language from specific *speech communities* (Gumperz, 1968).

In the second part of this paper, we apply basic deep NLP methods to building predictive models for abusive language on three different datasets. Through qualitative reflection on the data, training process, and results, we articulate the specific limitations of common methods, as well as the future directions, of deep learning methodology for addressing concerns about abusive language.

## 4 Experiments

The nascent research cluster around NLP and abusive language constitutes not just a 'speech community' but a *language community*, i.e., "an organization of people by their orientation to structural (formal) norms for denotational coding (whether explicit or implicit)" (Silverstein, 1996). The combination of linguistic ideologies described above is fully realized in the conventional experimental architecture of the *shared task*, in which multiple teams of researchers independently attempt to build systems with good classificatory performance by determining the true denotational meaning of utterances which, most commonly, have been excised from their interactional context.

For example, the tasks addressing abusive language typically have as their goal the determination of whether stand-alone utterances should be considered rude, offensive, or abusive. Training data is provided in the form of utterance-label pairs, where the label may be a binary value (i.e. abusive or not) or multi-class (for different categorical and/or ordinal levels of offensiveness). In order to explore these kinds of tasks directly, in this paper we chose to experiment with 3 datasets: the Kaggle Toxic Comment Classification Challenge[2], the shared task in the 1st Workshop on Trolling, Aggression and Cyberbullying (TRAC1)[3], and the StackOverflow dataset from the 2nd EMNLP Abusive Language Workshop.[4]

### 4.1 Data Description

The Kaggle Toxic Comment Classification dataset provides decontextualized Wikipedia "talk page" comments, each paired with multi-class labels on toxic behavior, judged by human raters; we emphasize that the dataset is *decontextualized* to indicate that additional information about each discursive interaction is not provided (but for a depiction of the organizational structure of their production, one may consult Geiger (2017)'s "ethnography of infrastructure" of Wikipedia). The labels of toxicity include 'toxic', 'severe toxic', 'obscene', 'threat', 'insult' and 'identity hate'. Because the other datasets we examine classify differing but related categories, it was necessary to combine these into one 'offensive' category to make comparisons across datasets possible (a common methodological decontextualization which elides available difference even at the level of the 'clean' dataset). 10.2% of the resulting dataset had an 'offensive' label. We split the data into training (150,571 observations), validation (6,000 observations) and holdout sets (3,000 observations).

The TRAC1 shared task dataset contains 15,000 stand-alone Facebook posts and comments in both Hindi and English unicode, each paired with human-rated multi-class labels, distinguishing "Overtly Aggressive", "Covertly Aggressive" and "Not Aggressive". There are separate English and Hindi subsets, and we used the English portion, which still contains significant amounts of Hindi-English code-switching (Verma, 1976). Again for comparison, it was necessary to group the first two categories together; in the resulting dataset, 58% of the comments are considered aggressive. We split the dataset into training (11,999 observations) and validation (3,001 observations) sets, and used the provided test set as our holdout set (601 observations).

The StackOverflow dataset is yet another collection of decontextualized comments, some of which are flagged by the users to be "Not Constructive or off topic", "Obsolete", "Other" (not the same as unflagged), "Rude or offensive", or "Too chatty". Notably, however, these flags are *provided by the site's users*; when a comment is flagged as "Rude or offensive", it is reportedly dynamically *removed from the website*, which makes this dataset's semantics different from the previous ones which were—as far as we can tell—labeled *post hoc* by independent raters. Instead, the StackOverflow data is a textual archive of *speech acts about speech acts*, or of *metapragmatic* utterances (Silverstein, 1993). They are traces of in-the-moment judgments that may have acted to spontaneously eliminate the judged utterance from a discourse.

The total number of flagged comments is

[2]https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge

[3]https://sites.google.com/view/trac1/home

[4]https://sites.google.com/view/alw2018/resources/stackoverflow-dataset

525,085, of which 57,841 are "Rude or offensive" (and thus were dynamically removed as per the above). In addition, there are 15 million comments that are not flagged to be undesirable in any way. We joined a sample of 1 million of the unflagged comments and the flagged comments, but considered only the flag "Rude or offensive" (the rest are grouped with unflagged). We used this dataset, which has 3.8% comments flagged as "Rude or offensive", for training and testing. We split this dataset into training (1,516,085), validation (6,000) and holdout (3,000) sets.

Out of the three datasets, both StackOverflow and Kaggle have a significant class imbalance, which is more significant for the StackOverflow set (3.8% offensive) than Kaggle (10.2% offensive).

### 4.2 Methods and Results

While some earlier research in the classification of abusive language used feature-based classification techniques such as support vector machines (Warner and Hirschberg, 2012), we were interested in evaluating deep learning methods comparable to work such as Founta et al. (2018). We implemented two neural network architectures widely used in text classification: a convolutional neural network (CNN) and a recurrent neural network using Bidirectional Gated Recurrent Units (RNN Bi-GRU).

**CNN Model** Convolutional neural networks (CNNs), originally popularized in the context of computer vision for recognition tasks (Le Cun et al., 1990), can be applied to sequences of word embeddings in a similar manner to how they are applied to bitmap images, and have been shown to perform well in some abusive language detection tasks (Park and Fung, 2017). Although CNNs are unlikely to capture longer-term sequential relations in the manner of the recurrent neural networks discussed below, they can plausibly capture local patterns of features, and offensive speech can often be detected by local features such as swear words/phrases and racial slurs.

We implemented a vanilla CNN using Keras (Chollet et al., 2015). The input is tokenized into words, and converted into 300-dimensional word embedding vectors using 1 million word vectors trained on Wikipedia using the Fasttext classifier (Joulin et al., 2016).[5] We set a maximum length of 100 tokens per input, and a vocabulary size of 30,000. The input layer is then fed into 2 convolutional layers (of kernel size 1*5) each followed by a max-pooling layer. This is followed by 2 dense layers (dimensions 128 and 64) and finally the output layer. We trained the model using the Adagrad optimizer (Duchi et al., 2011), using a batch size of 512 and 10 maximum epochs with early stopping.

**RNN Model** Recurrent neural networks are widely used in NLP tasks (e.g. Pavlopoulos et al. (2017)) because they are good at capturing longer-term sequential patterns in text. We used the RNN variant known as Bidirectional GRU (Chung et al., 2014; Cho et al., 2014); GRUs are recurrent units with both an update gate and a reset gate that aim to solve the "vanishing gradient" problem of vanilla RNN units.

We implemented the Bi-GRU model using Keras. The input layer is the same word embedding layer as the CNN model, which is fed into a 80-unit Bi-GRU layer, followed by a pooling layer concatenating features from an average and a max-pooling operation. This is then fed into the final output dense layer. We trained the model using the Adam optimizer (Kingma and Ba, 2014) and a dropout rate of 0.2, using a batch size of 512 and a maximum of 10 epochs with early stopping.

### 4.3 Results

| | **Offensive** | | **Normal** | | |
|---|---|---|---|---|---|
| Model | Prec | Recall | Prec | Recall | $F_1$ (micro) |
| **Kaggle Toxic** (327 offensive, 2673 normal) | | | | | |
| CNN | .74 | .76 | .97 | .97 | .86 |
| GRU | **.83** | **.76** | .97 | .98 | .89 |
| **TRAC1 Trolling** (354 offensive, 247 normal) | | | | | |
| CNN | .77 | 73 | .64 | .70 | .71 |
| GRU | .75 | .85 | .73 | .59 | .73 |
| **StackOverflow** (114 offensive, 2886 normal) | | | | | |
| CNN | .56 | .19 | .97 | .99 | .68 |
| GRU | **.59** | **.22** | .97 | .99 | .69 |

Table 1: Results on test sets of three data sources using two architectures. The numbers next to the data sources shows the size of each class in the test set. Hyperparameters were manually tuned using the validation sets. We calculated the micro-averaged F1 score because of the varied class imbalance in the datasets.

Our results show that the two architectures performed similarly, but there were large differences across the three datasets (see Table 1). The Kaggle dataset has the best results in terms of micro-averaged F1 score, with very high precision and

[5] https://fasttext.cc/docs/en/english-vectors.html

recall for the "normal" class and around 0.8 precision and recall for the "offensive" class. The TRAC1 dataset had a lower micro-averaged F1 score, but the performance on the two classes are more balanced than the Kaggle model. The StackOverflow dataset has the lowest micro-averaged F1 and the most unbalanced results between the two classes: high precision and recall for the non-offensive class, low precision and even lower recall (0.2) for the "offensive" class.

We argue that the large differences among the three datasets using the same architectures cannot be explained by differences in class imbalance; both Kaggle and StackOverflow have heavy class imbalance, yet the Kaggle model did much better on the offensive class (results highlighted in bold in Table1). Why, then, did the models perform so poorly on detecting offensive comments on StackOverflow?[6] Looking at the model predictions, we found that the predictions given by the GRU and the CNN models are highly correlated (Chi-squared = 1009.9, p $<$ 2.2e-16).[7] The mediocre precision on the "offensive class" is mainly caused by the fact that StackOverflow users don't always flag offensive comments, i.e., most of the false positives (where positive is a classification of 'offensive') should arguably be true positives. There are 23 comments that are predicted to be offensive by both models but don't receive an 'offensive' flag in the data. Out of the 23, two comments are indeed not (overtly or covertly) 'offensive': "`close(f)`–> `f.close()`"; "fuck bro !!! how the fuck didnt i see that , jesus !! thanksssssss !!!!!!". Among the rest, most are overtly offensive but not flagged, e.g. "dude can you answer the question or not? if not stop wasting my time"; "teach him instead of being a dick.". A few can be considered offensive in particular contexts or by certain users, e.g. "jesus christ! what're you doing?"; "don't. migrate. crap."; "no shit sherlock". This implies that the models would have had a higher precision if the gold standard was provided by annotators who judge every comment in the dataset. In this case, the pragmatic context of labeling matters.

The even lower recall, on the other hand, reveals a genuine limitation of the models and of the dataset. Again, the two models agree highly. Out of the 355 comments that are flagged as "offensive" by StackOverflow users, the majority (75%) are considered not offensive by both models (i.e. they are false negatives). 22 comments (6%) are identified as offensive by only one model, and only 52 comments (15%) are correctely labeled as offensive by both models. Why is the recall so low? To investigate, we sampled 100 of the false negatives and asked three human raters to determine whether these comments are offensive. Only 11 were considered offensive by at least two out of three raters even though they are flagged as "rude or offensive" by StackOverflow users. Here are some examples of comments flagged as offensive but *not* considered offensive by a majority of raters:

- *please post *code,* not screenshots.*
- *did not get you? where in the query that you have provided should i add this?*
- *the phrase is "want to."*
- *no testing!!!! i would prefer no coding*
- *you sir, deserve an unlimited amount of upvotes for that comment*

While these comments' lexical 'surface' content is unlikely to be considered offensive by our classifier, they can potentially be considered offensive in their pragmatic implicatures (Levinson, 1983), which can only be recovered or enriched given the context of the interaction and/or the broader context of conventions and norms in the StackOverflow forum.

Such context-dependent offensive comments appear to account for the majority of the false negatives in the StackOverflow results; this pattern is much less obvious in the Kaggle results. Unlike the StackOverflow dataset, the Kaggle dataset was constructed by showing annotators stand-alone comments. Therefore, the interactional context of those comments was not overtly considered during the rating, although it is likely that raters would sometimes imagine or "accommodate" context (Tian and Breheny, 2016). An analysis of the false negatives show that while a few comments likely require contextual enrichment (i.e., in the

[6] In this section we focus on predictions of offensive comments in the StackOverflow dataset, and compare it with results of Kaggle. Because of the heavy presence of Hindi and English code switching in the TRAC1 data, we did not perform an error analysis for this dataset. For in-depth discussions, please see the TRAC1 proceedings at https://sites.google.com/view/trac1/accepted-papers.

[7] We looked at predictions of both the validation set and the holdout set in order to have more samples to form a better understanding of the models.

referentialist ideology, they are "implicitly" offensive), the majority of the errors are due to unconventional ways of spelling, a known problem already being tackled by previous researchers who convincingly argue for character-level as opposed to word-level approaches (Mehdad and Tetreault, 2016).

To sum up, we saw that neural network models with different architectures (CNN and Bi-GRU) performed similarly and have the potential of reliable abusive/offensive language detection when the offensiveness is signaled and/or classified via expressions in the text-artifact itself (supported by the Kaggle results). However, when the offensiveness is marked in a context-dependent way, current neural network methods perform poorly; this is not necessarily because neural networks cannot be used to model context, but because the available datasets on abusive language detection do not provide this context. This is manifested in the poor performance of neural models on the StackOverflow data: the context-dependency of offensiveness results in low recall, and the inconsistency of user-generated flagging results in low precision. Because the flags are provided by users who have seen the entire interaction, many comments are considered offensive in context but not offensive when standing alone. By contrast, Kaggle and TRAC1 are labeled by independent annotators who did not participate or observe the full interaction.

## 5 Conclusions and Future Directions

In this paper, we have attempted to provide a quantitative justification for a qualitative perspective: namely, that theories of pragmatics (such as the primacy of context in the dynamic construction of meaning (Levinson, 1983)) and of metapragmatics (e.g. the fundamental reflexivity of interactional speech at various semiotic levels (Agha, 2007)) should take on a greater role in the classification of abusive language in NLP research.

Our experiments using common neural network architectures on text classification show promising performance when the offensiveness/ abusiveness is signalled within a single utterance, but give poor performance when the offensiveness require contextual enrichment. This is a limitation of popular abusive language detection tasks. For future work, we would propose to investigate the modeling of not just stand-alone utterances and their labels, but the affective and interactional dynamics of online communication.

In the case of StackOverflow, we suggest that a serious approach to tackling the problem of abusive language on the site would likely want to take advantage of the site's periodic data dumps, which provide millions of user questions, answers, votes, and favorites (Anderson et al., 2012). However, the dynamic removal of flagged material from the site poses some serious methodological issues, and the question of how to incorporate this vast relational data into neural network classifier architectures is another challenge, which we speculate will involve embeddings of networks of interactions as in Hamilton et al. (2018).

Finally, as a longer-term goal for the study of abusive language in online communities, we believe that it is quite promising that some researchers have implicitly or explicitly moved towards the notion of a *speech community*, in which actors in different social spaces may possess differing norms for appropriate behavior (Saleem et al., 2017). However, we argue that it will ultimately be necessary to attend to those theorists emphasizing so-called *communities of practice* (Holmes and Meyerhoff, 1999), a perspective which brings to the fore the embodiment of communities in *practical action* (of which language is only a part); to consider the role of conflict as well as consensus; to see identity as more than just a static set of categories; and to more seriously take into account the participants' understanding of their own practices (Bucholtz, 1999).

## 6 Acknowledgments

Many thanks to Ye Tian and Ioannis Douratsos for their assistance and suggestions; thanks also to Ana-Maria Popescu and Gideon Mann for their remarks. We would also like to thank the anonymous reviewers for their comments, critiques, and suggestions. Any opinions, findings, conclusions, or recommendations expressed here are those of the authors and do not necessarily reflect the view of their institutions.

## References

Asif Agha. 2007. *Language and Social Relations*, 1 edition edition. Cambridge University Press, Cambridge ; New York.

Ashton Anderson, Daniel Huttenlocher, Jon Kleinberg, and Jure Leskovec. 2012. Discovering Value from Community Activity on Focused Question Answering Sites: A Case Study of Stack Overflow. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '12, pages 850–858, New York, NY, USA. ACM.

J.L. Austin. 1962. *How to Do Things with Words*. Oxford University Press.

Richard Bauman and Charles L. Briggs. 1990. Poetics and Performance as Critical Perspectives on Language and Social Life. *Annual Review of Anthropology*, 19:59–88.

Alexander Brown. 2017a. What is hate speech? Part 1: The Myth of Hate. *Law and Philosophy*, 36(4):419–468.

Alexander Brown. 2017b. What is Hate Speech? Part 2: Family Resemblances. *Law and Philosophy*, 36(5):561–613.

Penelope Brown and Stephen C. Levinson. 1987. *Politeness: Some Universals in Language Usage*. Cambridge University Press.

Mary Bucholtz. 1999. "Why Be Normal?": Language and Identity Practices in a Community of Nerd Girls. *Language in Society*, 28(2):203–223.

Judith Butler. 1997. *Excitable Speech: A Politics of the Performative*. Routledge, New York.

Chaplinsky vs. New Hampshire. 1942. 315 U.S. 568.

Eugene Charniak. 1996. *Statistical Language Learning*. A Bradford Book, Cambridge, Mass.

Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*.

François Chollet et al. 2015. Keras. `https://keras.io`.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

Jonathan Culpeper. 1996. Towards an anatomy of impoliteness. *Journal of Pragmatics*, 25(3):349–367.

Mary Douglas. 1966. *Purity and Danger: An Analysis of the Concepts of Pollution and Taboo*. Routledge, London.

John Duchi, Elad Hazan, and Yoram Singer. 2011. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(Jul):2121–2159.

Alessandro Duranti. 1997. *Linguistic Anthropology*. Cambridge University Press, New York.

Antigoni-Maria Founta, Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Athena Vakali, and Ilias Leontiadis. 2018. A Unified Deep Learning Architecture for Abuse Detection. *arXiv:1802.00385 [cs]*. ArXiv: 1802.00385.

R. Stuart Geiger. 2017. Beyond opening up the black box: Investigating the role of algorithmic systems in Wikipedian organizational culture. *Big Data & Society*, 4(2):2053951717730735.

Erving Goffman. 1967. *Interaction Ritual: Essays on Face to Face Behaviour*. Penguin, Harmondsworth.

Paul Grice. 1975. Logic and Conversation. In P. Cole and N.L. Morgan, editors, *Syntax and Semantics, vol. 3: Speech Acts*, pages 41–58. Academic Press, New York.

John Gumperz. 1968. The Speech Community. In *International Encyclopedia of the Social Sciences*, pages 381–386. Macmillan, New York.

John J. Gumperz. 1982. *Discourse Strategies*. Cambridge University Press, Cambridge.

William L. Hamilton, Payal Bajaj, Marinka Zitnik, Dan Jurafsky, and Jure Leskovec. 2018. Querying Complex Networks in Vector Space. *arXiv:1806.01445 [cs, stat]*. ArXiv: 1806.01445.

Jane H. Hill. 2008. *The Everyday Language of White Racism*, 1 edition edition. Wiley-Blackwell, Chichester, U.K. ; Malden, MA.

Janet Holmes and Miriam Meyerhoff. 1999. The Community of Practice: Theories and Methodologies in Language and Gender Research. *Language in Society*, 28(2):173–183.

Jennifer Hornsby and Rae Langton. 1998. Free Speech and Illocution. *Legal Theory*, 4(1):21–37.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of Tricks for Efficient Text Classification. *arXiv:1607.01759 [cs]*. ArXiv: 1607.01759.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

William Labov. 1972. *Sociolinguistic Patterns*. University of Pennsylvania Press, Philadelphia.

Y. Le Cun, B. Boser, J. S. Denker, R. E. Howard, W. Habbard, L. D. Jackel, and D. Henderson. 1990. Advances in Neural Information Processing Systems 2. pages 396–404. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Michael Lempert. 2012. Implicitness. In Christina Bratt Paulston, Scott F. Kiesling, and Elizabeth S. Rangel, editors, *The Handbook of Intercultural Discourse and Communication*. John Wiley & Sons.

Stephen C. Levinson. 1983. *Pragmatics*. Cambridge University Press, Cambridge Cambridgeshire ; New York.

Christopher D. Manning. 2015. Computational Linguistics and Deep Learning. *Computational Linguistics*, 41(4):701–707.

Mari J. Matsuda. 1989. Public Response to Racist Speech: Considering the Victim's Story. *Michigan Law Review*, 87(8):2320–2381.

Yashar Mehdad and Joel Tetreault. 2016. Do Characters Abuse More Than Words? *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 299–303.

Sara Mills. 2003. *Gender and Politeness*. Cambridge University Press, Cambridge ; New York.

John Copeland Nagle. 2009. The Idea of Pollution. *UC Davis Law Review*, 43(1):1–78.

Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive Language Detection in Online User Content. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, pages 145–153, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.

Ji Ho Park and Pascale Fung. 2017. One-step and two-step classification for abusive language detection on twitter. *arXiv preprint arXiv:1706.01206*.

John Pavlopoulos, Prodromos Malakasiotis, and Ion Androutsopoulos. 2017. Deep learning for user comment moderation. *arXiv preprint arXiv:1705.09993*.

Fernando Pereira. 2000. Formal grammar and information theory: together again? *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 358(1769):1239–1253.

Björn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. 2017. Measuring the Reliability of Hate Speech Annotations: The Case of the European Refugee Crisis. *arXiv:1701.08118 [cs]*. ArXiv: 1701.08118.

Haji Mohammad Saleem, Kelly P. Dillon, Susan Benesch, and Derek Ruths. 2017. A Web of Hate: Tackling Hateful Speech in Online Social Spaces.

Anna Schmidt and Michael Wiegand. 2017. A Survey on Hate Speech Detection using Natural Language Processing. *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10.

Lisa H. Schwartzman. 2002. Hate Speech, Illocution, and Social Context: A Critique of Judith Butler. *Journal of Social Philosophy*, 33(3):421–441.

John Searle. 1969. *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press.

Michael Silverstein. 1993. Metapragmatic Discourse and Metapragmatic Function. In John Lucy, editor, *Reflexive Language: Reported Speech and Metapragmatics*, pages 33–58. Cambridge University Press, Cambridge.

Michael Silverstein. 1996. Encountering Language and Languages of Encounter in North American Ethnohistory. *Journal of Linguistic Anthropology*, 6(2):126–144.

Caroline Sinders and Freddy Martinez. 2018. Online Monitoring of the Alt-Right. The Circle of HOPE, New York City, 27th July 2018.

Ellen Spertus. 1997. Smokey: Automatic Recognition of Hostile Messages. In *Proceedings of the Fourteenth National Conference on Artificial Intelligence and Ninth Conference on Innovative Applications of Artificial Intelligence*, AAAI'97/IAAI'97, pages 1058–1065, Providence, Rhode Island. AAAI Press.

Ye Tian and Richard Breheny. 2016. Dynamic pragmatic view of negation processing. In *Negation and polarity: Experimental perspectives*, pages 21–43. Springer.

S. K. Verma. 1976. Code-switching: Hindi-English. *Lingua*, 38(2):153–165.

William Warner and Julia Hirschberg. 2012. Detecting Hate Speech on the World Wide Web. In *Proceedings of the Second Workshop on Language in Social Media*, LSM '12, pages 19–26, Stroudsburg, PA, USA. Association for Computational Linguistics.

Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding abuse: A typology of abusive language detection subtasks. *CoRR*, abs/1705.09899.

Ludwig Wittgenstein. 1953. *Philosophical investigations*. Macmillan Publishing Company.

Kathryn A. Woolard. 1998. Introduction: Language Ideology as a Field of Inquiry. In Bambi B. Schieffelin, Kathryn A. Woolard, and Paul V. Kroskrity, editors, *Language Ideologies: Practice and Theory*, pages 3–32. Oxford University Press, U.S.A., New York.

Kathryn A. Woolard and Bambi B. Schieffelin. 1994. Language Ideology. *Annual Review of Anthropology*, 23(1):55–82.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2016. Ex Machina: Personal Attacks Seen at Scale. *arXiv:1610.08914 [cs]*. ArXiv: 1610.08914.