I Beg to Differ:
A study of constructive disagreement in online conversations

Christine De Kock and Andreas Vlachos
University of Cambridge
Department of Computer Science and Technology
{cd700,av308}@cam.ac.uk

Abstract

Disagreements are pervasive in human communication. In this paper we investigate what makes disagreement constructive. To this end, we construct WikiDisputes, a corpus of 7 425 Wikipedia Talk page conversations that contain content disputes, and define the task of predicting whether disagreements will be escalated to mediation by a moderator. We evaluate feature-based models with linguistic markers from previous work, and demonstrate that their performance is improved by using features that capture changes in linguistic markers throughout the conversations, as opposed to averaged values. We develop a variety of neural models and show that taking into account the structure of the conversation improves predictive accuracy, exceeding that of feature-based models. We assess our best neural model in terms of both predictive accuracy and uncertainty by evaluating its behaviour when it is only exposed to the beginning of the conversation, finding that model accuracy improves and uncertainty reduces as models are exposed to more information.

1 Introduction

Disagreements online are a familiar occurrence for any internet user. While they are often perceived as a negative phenomenon, disagreements can be useful; as is illustrated in Figure 1, disagreements can lead to an improved understanding of a topic by introducing and evaluating different perspectives. Research on online disagreements has focused on mostly negative aspects such as trolling (Cheng et al., 2017), hate speech (Waseem and Hovy, 2016), harassment (Yin et al., 2009) and personal attacks (Wulczyn et al., 2017). Recent works by Zhang et al. (2018) and Chang and Danescu-Niculescu-Mizil (2019) study conversations on Wikipedia talk pages that start out as civil but derail into personal attacks, using linguistic markers and deep learning approaches, respectively.

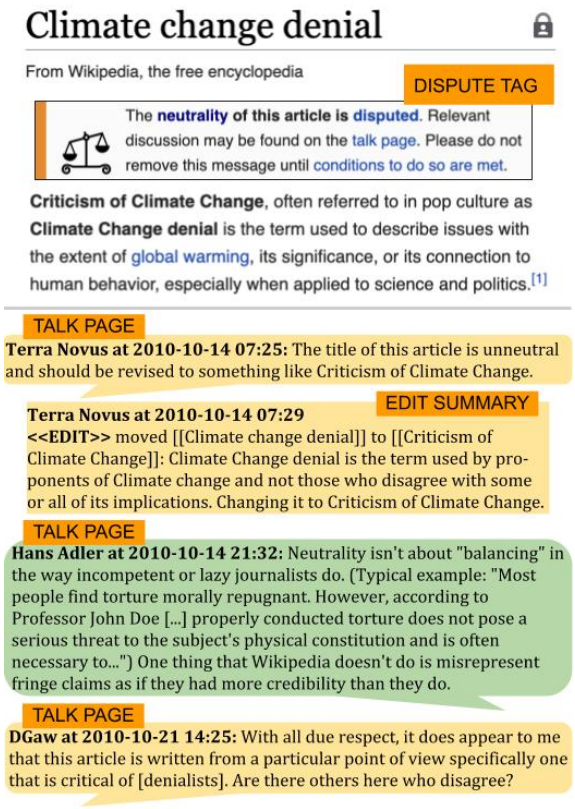

Figure 1: An example of a dispute on Wikipedia, showing the dispute tag on the article, talk page posts, and summaries of edits occurring during the discussion.

An alternative approach is to study good faith disagreement through debates on online platforms such as ChangeMyView (Tan et al., 2016), debate.org (Durmus and Cardie, 2019) and Kialo (Boschi et al., 2019), or formal Oxford-style debates (Zhang et al., 2016a). While this has benefits, such as readily available annotation of stances and indication of winning and losing sides, it does not mirror the way people naturally converse. For example, in formal debates, temporal and structural

Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, pages 2017–2027
April 19 - 23, 2021.

constraints are imposed. Moreover, in Oxford-style debates, participants are not motivated to come to a consensus but rather to persuade an audience of a predefined stance (Zhang et al., 2016a). Finally, the task of detecting disagreement in conversations has been studied by a number of authors, e.g. Wang and Cardie (2014) and Rosenthal and McKeown (2015), without attempting to analyze it further.

We are interested instead in *constructive disagreement in an uncoerced setting*. To this end, we present `WikiDisputes`[1]; a corpus of 7 425 disagreements (totalling 99 907 utterances) mined from Wikipedia Talk pages. To construct it, we map dispute tags in the platform's edit history (shown in Figure 1) to conversations in WikiConv (Hua et al., 2018) to locate conversations that relate to content disputes. Observing that conversations are often conducted in the Talk pages and the edit summaries simultaneously, as illustrated in Figure 1, we augment the WikiConv conversations with edit summaries that occur concurrently. To investigate the factors that make a disagreement constructive, we define the task of predicting whether a dispute is eventually considered resolved by its participants, or it was escalated by them to mediation by a moderator, and therefore considered unconstructive.

We evaluate both feature-based and neural models for this task. Our findings indicate that balancing hedging and certainty plays an important role in constructive disagreements. We find that incorporating edit summaries (which have been ignored in previous work on Wikipedia discussions) improves model performance on predicting escalation. We further find that including information about the conversation structure aids model performance. We observe this in two ways. Firstly, we include gradient features, which capture changes in linguistic markers throughout the conversations as opposed to averaged values. Secondly, we experiment with adding sequential and hierarchical conversation structure to our neural models, finding that a Hierarchical Attention Network (Yang et al., 2016) provides the best performance on our task.

We further evaluate this model in terms of its predictive accuracy when exposed to the beginnings of the conversation as opposed to completed conversations, as well as its uncertainty (more details in Section 6.4). Our results indicate that model performance is reduced from a PR-AUC of 0.29 to 0.19 when exposed to only the first half of the conversation. However, this reduced model still outperforms toxicity and sentiment models predicting on full conversations. Model uncertainty decreases roughly linearly as the model is exposed to more information. We conduct a qualitative analysis of the points in the conversation where the model changes its prediction on constructiveness, finding that some markers from our feature-based models also seem to have been picked up by the neural model.

2 Dispute resolution on Wikipedia

Wikipedia relies heavily on collaboration and discussion by editors to maintain content standards, using the platform's Talk pages (Wikipedia, 2020c). Wikipedia contributors add "dispute" templates to articles (Wikipedia, 2020a) referring to a content accuracy dispute or to a violation of the platform's neutral point of view (NPOV) policy (Wikipedia, 2020b). Adding such a template to the article creates a dispute tag on the article as shown in Figure 1. Variations of these tags allow contributors to flag specific sections or phrases which relate to the issue in question, or to add details of the dispute.

All edits to Wikipedia, including the aforementioned dispute tags, are retained in the site's edit history. Revision metadata contains a description of the edit (hereafter referred to as the edit summary). As observed in Figure 1, when such edits occur while a conversation is underway, editors use the "edit summaries" to clarify their intentions in the context of the conversation.

The Wikipedia dispute resolution policy stipulates that if contributors are unable to reach a consensus through Talk page discussion, they can request mediation by a community volunteer. A prerequisite for this escalation step is proof of recent, extensive discussion on a Talk page[2].

3 WikiDisputes

Our dataset consists of three facets: disagreements on Talk pages, edit summaries, and escalation labels. Wikipedia informs users that all contributions on article or Talk pages are published publicly, and provides channels for users to permanently remove personal information that they have shared accidentally or otherwise (Wikipedia, 2020d), thus we are able to release this data to the community. In

[1] `github.com/christinedekock11/wikidisputes`

[2] `https://en.wikipedia.org/wiki/Wikipedia:Dispute_resolution_noticeboard`

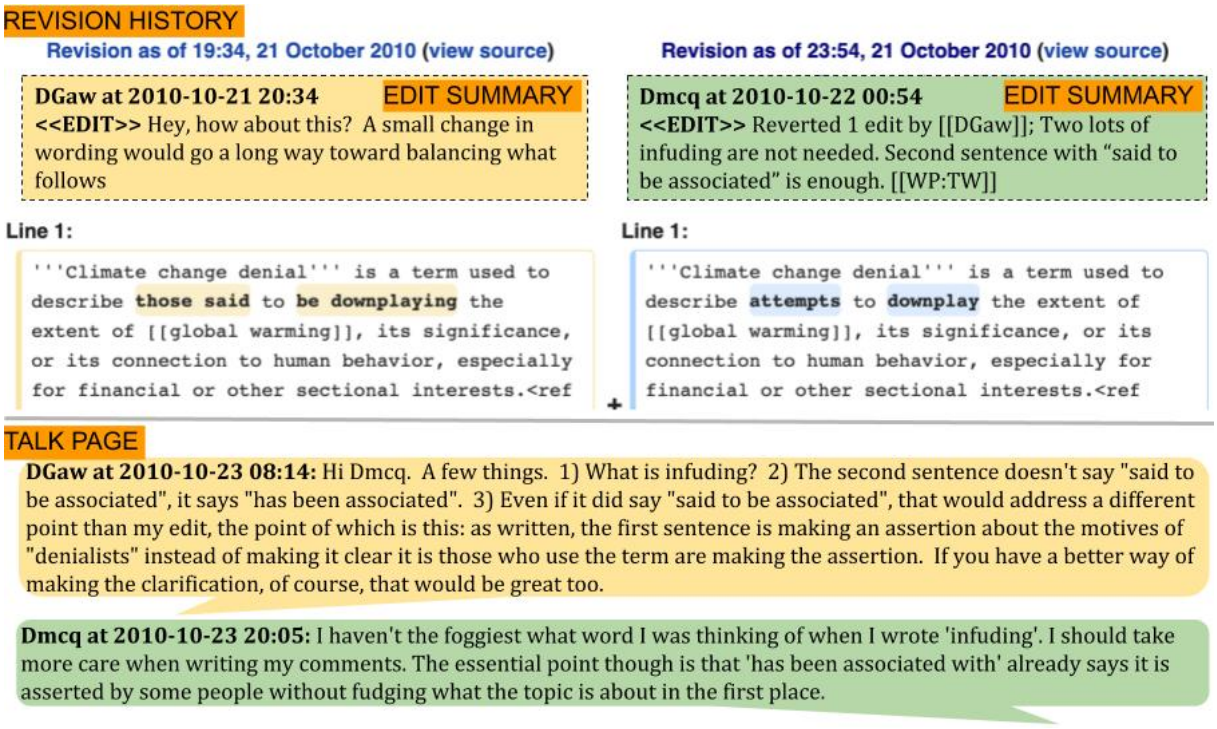

Figure 2: An illustration of the relationship between the revision history and Talk page conversations, using later utterances from the dispute in Figure 1. Two consecutive edits and their summaries are shown side by side.

the remainder of this section, we detail the process for obtaining the conversations, the edit summaries and the labels.

3.1 Finding disagreements

We use the Wikipedia revision history dump[3] from 1 July 2020 to find content disputes. We locate the addition of dispute templates in the article history using regular expressions that account for variations in the free text dispute templates. Using the revisions where dispute tags are added, we can also determine which user logged the dispute, the timestamp, and the article and section in dispute.

To find the related WikiConv conversation for each dispute tag, we select all conversations on the relevant article's Talk page in which the user who logged the dispute was involved. Of these candidates, the conversation closest in time to the timestamp of the dispute tag addition is selected. We filter the conversations extracted above using a number of heuristics in the pursuit of high quality samples of constructive disagreements. To ensure conversations of adequate length, we set a minimum conversation length of five utterances and 250 tokens. Based on an inspection of 100 conversations, we remove conversations of more than 50 utterances, as such conversations are often flame wars. We further include only conversations in which more than one user participated. Using this methodology, we obtain 7 425 content disputes.

Our manual inspection further indicates that the usage of the dispute tag is not entirely consistent across the platform; editors sometimes use the tag to report that they dispute some aspect of the text before discussing it, rather than to flag a currently occurring dispute in the Talk page. However, the former frequently results in a disagreement. We found that 88 out of 100 inspected examples were correctly identified disagreements, which we believe is an acceptably low error rate.

3.2 Collecting edit summaries

We show in Figure 2 two further posts from the climate dispute (continued from Figure 1), along with two edits that occurred during the course of the conversation. We note here that including the edit summaries as part of the conversation is important for understanding the conversation; without this, the reference to "infuding" in this example would be incomprehensible. We thus include all edit summaries written by users involved in the conversation, which are timestamped between the first and last utterance in the conversation.

[3] https://dumps.wikimedia.org/enwiki/

3.3 Escalation tags

For the task of inferring whether a dispute was escalated to mediation (and therefore Talk page discussion was unsuccessful), we scrape mediated cases from the Dispute Resolution Noticeboard archives and align them with WikiConv conversations.

Of the 2 520 archived mediation processes, 149 were closed as failures and 237 as successes, with the remaining 2 134 receiving "General closures". The last label is frequently assigned due to insufficient Talk page discussion. We only include accepted mediations in our dataset and thus ignore "General closures", resulting in 386 mediations.

We record the location of the disagreement to which the mediation relates, the usernames of participants and the timestamp. We use a similar alignment procedure to that described in Section 3.1 to find the relevant WikiConv conversations; except that we include the usernames of all participants listed in the mediation. During alignment, a simple match is recorded when there is exactly one conversation on the article Talk page involving all listed users (118 cases out of 386).

There are 111 cases in which no full matches occur; this sometimes happens when a participant changes their username or is participating anonymously. We manually inspect potential matches in this case. A further 67 cases have multiple matches; this happens when a participant changes their username, or is participating anonymously, or the Talk page has been archived; in this case we try to find the conversation most related to the mediation summary.

We use only the utterances which are timestamped from before the conversation was escalated and apply the same filters outlined in Section 3.1. Using this methodology, we are able to find 201 disagreements that are escalated to mediation. The complementary class label (not escalated) is assigned to the disagreements extracted using the dispute tags, for which no mediation was found.

Our classification task uses escalation as a proxy label for cases where Talk page discussion was not constructive. This relies on the assumption that non-escalated disagreements are more constructive than escalated disagreements, even though disagreements that were not escalated are not guaranteed to be successfully resolved; participants may simply not be aware of the escalation procedure, or not be willing to go through the extra effort of participating in mediation. To validate this, we annotated 35 samples from the dataset as to whether they represented constructive disagreements, and our annotations agreed with the escalation labels in 25 of these. Combined with the low estimates of antisocial behaviour on the site (Wulczyn et al., 2017), we draw the inference that the non-escalated class contains more constructive disagreements than the escalated ones. Throughout the remainder of this paper, we use these terms interchangeably.

4 Modelling constructive disagreement

4.1 Feature-based models

We develop our feature-based models using feature sets suggested in previous research on tasks related to conversational analysis. These include:

- **Politeness**: The politeness strategies from Zhang et al. (2018) as implemented in Convokit (Chang et al., 2020), which capture greetings, apologies, directness, and saying "please", etc.
- **Collaboration**: Conversation markers in collaborative discussions (Niculae and Danescu-Niculescu-Mizil, 2016), which capture the introduction and adoption of ideas, uncertainty and confidence terms, pronoun usage, and linguistic style accommodation.
- **Toxicity**: Toxicity and severe toxicity scores as estimated by the Perspective API (Wulczyn et al., 2017) and included in WikiConv (Hua et al., 2018).
- **Sentiment**: Positive and negative sentiment word counts, as per the lexicon of Liu et al. (2005) and implemented in Convokit (Chang et al., 2020).

For each feature, we calculate the average value throughout the conversation, as well as the gradient of a straight line fit of the feature value throughout the conversation. Our intuition for including this variation is that the outcome of a disagreement is likely to be influenced by a change in the language usage through the course of a conversation, which is not reflected by the mean. For instance, a disagreement might start out very politely and end impolitely, or the other way around, and would have the same mean. Logistic regression is used to infer linear relationships between the linguistic features and disagreement outcomes.

4.2 Neural models

Dialogue structure has been difficult to capture in feature based models due to sparsity. For this reason, we implement a number of neural models with increasing capacities for modelling conversation structure to assess its importance.

Averaged embeddings Our simplest variant averages the GloVe embeddings (Pennington et al., 2014) of all words in a conversation, and uses a fully connected layer for classification. It ignores both utterance hierarchy and word ordering and can be seen as a bag-of-word-embeddings approach.

LSTM Instead of averaging the GloVe embeddings, we use a bidirectional LSTM-based model (Hochreiter and Schmidhuber, 1997) to process the sequence of words in the conversation; ignoring utterance hierarchy but preserving word order.

HAN We use a Hierarchical Attention Network (HAN) (Yang et al., 2016) to model both word order and utterance hierarchy. HANs have recently been used to predict emotion in conversations (Ma et al., 2020) and are useful for capturing the structure of texts by building up utterance embeddings from word embeddings, and then constructing a conversation vector from these utterance embeddings. Given a sequence of words in an utterance, the words are first mapped to vectors through an embedding matrix (GloVe embeddings, in our case). A bidirectional LSTM layer (Hochreiter and Schmidhuber, 1997) is used to process the sequence of embeddings and calculate an embedding for each word in the context of the utterance. An attention mechanism (Bahdanau et al., 2015) is applied to these embeddings to find an utterance embedding. A similar procedure is followed to build up a conversation vector based on the utterances embeddings. This vector is then used to perform the classification. To incorporate edit summaries in a conversation, we prepend each edit summary with a special token (<EDIT>) to indicate its origin.

5 Experimental setup

An initial data analysis indicated that escalated disagreements have a median length of 16 utterances, compared to 11 for their non-escalated counterparts. While this may be an informative feature for classification, we are primarily interested in the effects of language on the conversation outcome, and therefore choose to control for this effect. We use *matching*, a technique developed for causal inference in observational studies (Rubin, 2007), also employed by Zhang et al. (2018) in the context of conversational modelling. We pair each escalated disagreement with a non-escalated disagreement of the same length in utterances, bearing in mind that the dataset is heavily imbalanced, with 7 425 non-escalated disagreements compared to 201 escalated disagreements. To retain an imbalance in the classes while conducting matching, we match every escalated disagreement with up to ten non-escalated disagreements (the actual number depending on availability) by randomly sampling without replacement from the non-escalated disagreements of the same length. Characteristics of the dataset after performing matching are shown in Table 1, with both classes having a median of 16 utterances now. Escalated disagreements have one more participant than non-escalated disagreements on average, and utterances in the escalated disagreements class are slightly shorter.

	Not escalated	**Escalated**
# Samples	1994	216
# Participants	2	3
# Utterances	16	16
Tokens per utt.	61	53

Table 1: Dataset statistics for the task of predicting escalation. Where applicable, median values are used.

We split the dataset for training as indicated in Table 2. Due to the class imbalance we use the area under the precision-recall curve (Davis and Goadrich, 2006) as metric for this task; however, for the sake of interpretability we also present the break-even F1 scores. We use a distribution-aware random class predictor as a random baseline.

Hyperparameters The logistic regression models are implemented in Scikit-Learn. We use a grid search to determine the best regularisation mode (L1 or L2) per model, evaluating C-values in [0.1,1,10,100]. The neural models are implemented in Keras, using Focal Loss (Lin et al., 2017) with Adam optimisation (Kingma and Ba, 2014) (learning rate=0.001) and Dropout ($p = 0.3$). For the HAN, we use bidirectional LSTM layers with 128 nodes in both the utterance and conversation encoders. For the LSTM model, we use only one such a layer.

Dataset	Escalated	Not escalated
Train	125	1411
Test	46	284
Validation	30	299

Table 2: Test set splits for predicting escalation.

Model	PR-AUC	F1
Baselines		
Random	0.121	0.128
Bag-of-words	0.213	0.239
Feature-based models		
Toxicity	0.140	0.125
Sentiment	0.150	0.055
Politeness	0.232	0.241
+ *gradients*	0.275	0.243
Collaboration	0.261	0.320
+ *gradients*	0.269	0.302
Politeness and collaboration	0.255	0.256
+ *gradients*	0.281	0.289
Neural models		
Averaged embeddings	0.243	0.256
LSTM	0.263	0.194
HAN	0.373	0.304
+ *edit summaries*	**0.400**	**0.333**

Table 3: Results of the escalation prediction task.

6 Results

Results are shown in Table 3. We note that generally, the neural network models perform better than the feature-based models. We discuss the PR-AUC scores of each model category below, noting that the scores from the F1 metric generally follow the same trends, but that PR-AUC is more robust in the face of imbalanced data. Furthermore, we investigate whether it is possible to predict the outcome from the beginning, how model uncertainty changes, or if there is often some inflection point in a conversation that changes the outcome.

6.1 Feature-based models

A number of our feature-based models outperforms the bag-of-words baseline, which in turn performs better than the random baseline. The toxicity and sentiment features do not perform better than the bag-of-words baseline, which indicates that these features by themselves do not capture constructiveness in disagreements. Toxicity scores were found by Zhang et al. (2016a) to be predictive of a conversation derailing, but in our case seemingly lead to the inference of spurious correlations. An explanation might be that comments which seem toxic out of context are in fact spirited discussion due to the degree of involvement of the participants; for instance, we have observed that contributors in some cases challenge others' credentials, which may seem rude in casual conversation, but can be useful in resolving a dispute.

Feature	Type	Coeff.
2nd person pronouns, $\bar{x}$	Both	+3.64
# hedging terms, $\bar{x}$	Both	-2.48
Greetings, $\bar{x}$	Politeness	+2.30
1st person pronouns, $\bar{x}$	Both	+1.91
Greetings, ∇	Politeness	-1.81
Deference, $\bar{x}$	Politeness	-1.44
3rd person pronouns, ∇	Collaboration	-1.38
# ideas adopted + certainty, $\bar{x}$	Collaboration	-1.23
Use of "by the way", $\bar{x}$	Politeness	-1.04
Certainty,∇	Collaboration	-0.92

Table 4: Ten features with the largest coefficients of the best feature-based model, which uses politeness and collaboration featuresets. Feature variations are the mean ($\bar{x}$) and gradient (∇). Positive weights are associated with the unconstructive class.

The best featuresets proposed in previous work are collaboration (Niculae and Danescu-Niculescu-Mizil, 2016), followed closely by politeness (Zhang et al., 2016a). Adding the gradient features we proposed improves performance for both, indicating that not only the presence of a marker is important, but also how its usage throughout a conversation changes. The best feature-based model is a combination of politeness and collaboration features, with a PR-AUC of 0.281 ($P < 0.05$, using a randomized permutation test).

The ten features with the largest coefficients from the combined model are shown in Table 4 in descending order or magnitude, along with their directions. Positive coefficients are associated with the escalated class, which indicates that a disagreement was not constructive.

Politeness markers such as deference and the use of "by the way" are found indicative of constructiveness, corroborating the findings of Zhang et al. (2016b). Greetings are associated with unconstructive disagreements on average, but an increase in greetings towards the end of a conversation is associated with constructiveness. An explanation of this might be that greetings can seem overly formal or indicative of tension early on, but if used later in a conversation, they indicate that new participants have entered the conversation or that more time is taken between replies. Indeed, longer gaps between replies (a feature in the collaboration featureset) are also associated with constructiveness.

The use of first and second person pronouns ("I" and "you") is associated with unconstructive disagreements, with the latter corroborating the findings of Zhang et al. (2018). This is also consistent

with psychotherapy research on disagreements in relationships (e.g. Gottman and Krokoff (1989)), which emphasises avoiding 'you'-messages which might be perceived as blameful.

Hedging is associated with constructive disagreements, which corroborates the findings of Zhang et al. (2016a). An intuitive understanding of this result is that using hedging terms shows that the speaker is more open to adjusting their opinion or compromising. The certainty gradient and the adoption of new ideas with certainty are also associated with constructiveness. Niculae and Danescu-Niculescu-Mizil (2016) found no significant correlation between either certainty or hedging and successful collaboration. We attribute the observed differences in feature associations to the fact that WikiDisputes are sourced from an uncoerced setting, where users feel more invested in the outcome of a disagreement and may need to balance the use of certainty and hedging to negotiate compromises. The setting of Niculae and Danescu-Niculescu-Mizil (2016) obligates volunteers to participate in a photo geolocation game, where it is unlikely that interlocutors are as invested in the task as collaborators working on a widely read encyclopaedia.

Words associated with either class, obtained through the bag-of-words model, are shown in Appendix A. An interesting observation from this list is that citing Wikipedia policy (which is indicated with WP) is associated with escalating conversations. This practice is sometimes referred to as "Wiki-Lawyering" within the community and can signal an unwillingness to compromise.

6.2 Neural network models

The results from our neural models illustrate that incorporating structure improves predictive accuracy on our task.

The model that averages over word embeddings performs the worst; a moderate increase in performance (2%) is gained from adding sequential word processing by way of an LSTM model. Adding utterance boundary information with HAN results in a much larger improvement (11%) over the LSTM. This indicates that, while it is helpful to observe the ordering of words in a conversation, a more critical component in the case of disagreements is how words are arranged in utterances.

The highest scoring model, which a PR-AUC of 0.40, results from including edit summaries in the conversations ($P < 0.05$, using a randomized permutation test). This provides support for our observation that understanding some conversations requires knowledge of the edits that occurred during the conversation. However, neglecting to insert the "EDIT" token (as explained in Section 4.2) results in the PR-AUC decreasing to 0.33, which indicates that utterances sourced from edit summaries require special processing and are not completely homogeneous in the conversation.

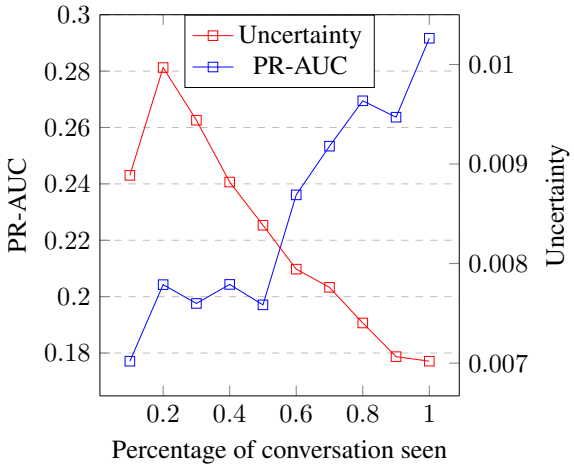

Figure 3: PR-AUC scores and uncertainty values of a HAN model when exposed to partial conversations.

Evaluating the predictions of this model, we observe that false positives often co-occur with heated debate. This is not unexpected, given that the positive class is contains interlocutors who feel strongly enough about their case to request mediation. The false negatives, on the other hand, contain a number of cases where it seemed as though a disagreement had been resolved, but then the argument progresses as further edits are made.

6.3 Early estimation of outcome

We are interested in how model predictions change throughout the conversation; whether it is possible to predict the outcome from the beginning, or if there is often some inflection point in a conversation that changes the outcome. Predicting the outcome of a conversation based on only a few utterances means that the model has less information for its inference; however, if models can perform well under such conditions, early intervention by a mediator could be recommended.

To evaluate model performance in such conditions, we split each disagreement into 10 buckets, chronologically (using only conversations of more than 10 utterances). We evaluate models trained on

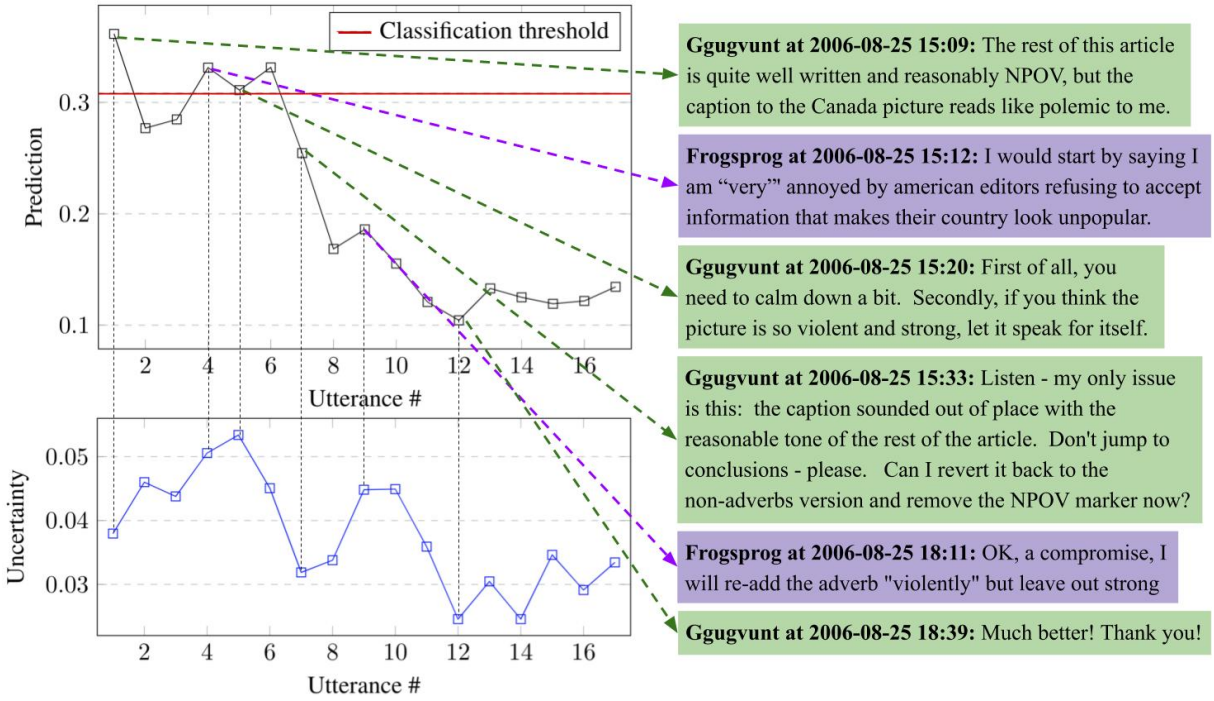

Figure 4: Model predictions and uncertainty throughout the course of a conversation in our dataset. The classification threshold shown is tuned on the validation set.

full conversations on these subsets to observe the change in model performance. These results are shown in Figure 3 in blue. Model performance increases from 0.198 at the midway point (0.5 bucket) to 0.292 when the full conversation has been observed, indicating that signals later in conversations are often important for predicting outcome. However, the neural model's performance when predicting on half the conversation only is still better than the toxicity and sentiment models on the complete conversations.

6.4 Uncertainty estimation

Given that this early estimation exposes the model to less information, we are interested in whether this impacts model uncertainty. We employ Monte Carlo dropout (Gal and Ghahramani, 2016) to this end. This approach calls for applying dropout before every layer of weights during both training and prediction, allowing us to sample from the distribution of predicted values for each input. Each input is evaluated multiple times (in our case, $N = 30$) and the mean and standard deviation of these predictions represent the prediction and its uncertainty.

Our results are shown in Figure 3 (in red). We note that uncertainty initially increases and then decreases monotonically. The initial increase in uncertainty could be due to a contrary position being introduced in the second utterance, in response to the introductory comment (as occurs in Figure 4). From there, the dispute is either resolved or eventually escalated, and uncertainty decreases almost linearly as the model is exposed to new data.

6.5 Identifying inflection points

Having ascertained that model accuracy improves and uncertainty decreases as a model is exposed to more information, we are interested in factors that cause the predicted class to change, and how the model predictions relate to the feature-based model coefficients. Although methods for interpreting neural network predictions exist (Ribeiro et al., 2016, 2018), these are not easily extendable to process inter-utterance dependencies as observed with the HAN. We instead analyse a conversation from our dataset to observe inflection points and what may have caused them. We show HAN model predictions and uncertainty values in Figure 4.

The HAN model initially predicts escalation, with a relatively low uncertainty value. The conversation remains confrontational for the first seven utterances with a high escalation score. There are two "I" pronouns in the second utterance and two "you" pronouns in the third utterance, which were also associated with unconstructiveness in the feature-based models. We note the use of politeness cues

("please" and "thank you" in the fourth and sixth utterances in this example), which reduces the score for escalation. The uncertainty and prediction values increase when one user proposes a compromise in utterance 9, and then decreases again in utterance 12 when the other user accepts the compromise. This feature was not explicitly modelled for in our feature-based models.

7 Conclusion

In a discussion of online disagreements by Graham (2008) (which Wikipedia references in its guidelines for resolving disputes), the author remarks that an increase in disagreements through widespread online interaction has the potential to create a surge in anger among internet users. On the other hand, as illustrated in our data, disagreements can sometimes be constructive. The success of Wikipedia shows the benefits of integrating multiple perspectives. As Hahn (2020) states, "Arguing things through is at the center of our attempts to come to accurate beliefs about the world [and] decide on a best course of action." For this reason, it is important to understand why disagreement can sometimes be constructive, and in other cases leads to conversational failure.

In this work, we investigated constructive disagreements from an NLP perspective. We have proposed a dataset of disagreements online and defined the task of predicting escalation as a proxy label for cases where disagreements were unconstructive. We analysed features that are associated with either class, and drew parallels with existing work. Using neural network models, we investigated the effect of modeling conversation structure and found that adding utterance hierarchy lead to an increase in performance. Finally, we validated our neural models by evaluating their performance and uncertainty when exposed to partial conversations. Our insights on constructive disagreements are not limited to Wikipedia and would be transferable to disagreements on other platforms. Additionally, our finding that edit summaries are an informative part of Talk page conversations should be useful for researchers who work on Wikipedia Talk pages more generally.

Acknowledgements

Christine De Kock is supported by scholarships from Huawei and the Oppenheimer Memorial Trust. Andreas Vlachos is supported the EPSRC grant no. EP/T023414/1: Opening Up Minds.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015*.

Gioia Boschi, Anthony P Young, Sagar Joglekar, Chiara Cammarota, and Nishanth Sastry. 2019. Having the last word: Understanding how to sample discussions online. *arXiv preprint arXiv:1906.04148*.

Jonathan P. Chang, Caleb Chiam, Liye Fu, Andrew Wang, Justine Zhang, and Cristian Danescu-Niculescu-Mizil. 2020. ConvoKit: A toolkit for the analysis of conversations. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 57–60, 1st virtual meeting. Association for Computational Linguistics.

Jonathan P. Chang and Cristian Danescu-Niculescu-Mizil. 2019. Trouble on the horizon: Forecasting the derailment of online conversations as they develop. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4743–4754, Hong Kong, China. Association for Computational Linguistics.

Justin Cheng, Michael Bernstein, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2017. Anyone can become a troll: Causes of trolling behavior in online discussions. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*, pages 1217–1230.

Jesse Davis and Mark Goadrich. 2006. The relationship between Precision-Recall and ROC curves. In *Proceedings of the 23rd international conference on Machine learning*, pages 233–240.

Esin Durmus and Claire Cardie. 2019. A corpus for modeling user and language effects in argumentation on online debating. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 602–607, Florence, Italy. Association for Computational Linguistics.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059.

John M Gottman and Lowell J Krokoff. 1989. Marital interaction and satisfaction: A longitudinal view. *Journal of consulting and clinical psychology*, 57(1):47.

Paul Graham. 2008. How to disagree. http://www.paulgraham.com/disagree.html.

Ulrike Hahn. 2020. Argument quality in real world argumentation. *Trends in Cognitive Sciences*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Yiqing Hua, Cristian Danescu-Niculescu-Mizil, Dario Taraborelli, Nithum Thain, Jeffery Sorensen, and Lucas Dixon. 2018. WikiConv: A corpus of the complete conversational history of a large online collaborative community. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2818–2823, Brussels, Belgium. Association for Computational Linguistics.

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

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.

Bing Liu, Minqing Hu, and Junsheng Cheng. 2005. Opinion observer: Analyzing and comparing opinions on the web. In *Proceedings of the 14th international conference on World Wide Web*, pages 342–351.

Hui Ma, Jian Wang, Lingfei Qian, and Hongfei Lin. 2020. HAN-ReGRU: hierarchical attention network with residual gated recurrent unit for emotion recognition in conversation. *Neural Computing and Applications*, pages 1–19.

Vlad Niculae and Cristian Danescu-Niculescu-Mizil. 2016. Conversational markers of constructive discussions. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 568–578, San Diego, California. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Anchors: High-precision model-agnostic explanations. In *AAAI*, volume 18, pages 1527–1535.

Sara Rosenthal and Kathy McKeown. 2015. I couldn't agree more: The role of conversational structure in agreement and disagreement detection in online discussions. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 168–177, Prague, Czech Republic. Association for Computational Linguistics.

Donald B Rubin. 2007. The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. *Statistics in medicine*, 26(1):20–36.

Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2016. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of the 25th international conference on world wide web*, pages 613–624. International World Wide Web Conferences Steering Committee.

Lu Wang and Claire Cardie. 2014. A piece of my mind: A sentiment analysis approach for online dispute detection. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 693–699, Baltimore, Maryland. Association for Computational Linguistics.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California. Association for Computational Linguistics.

Wikipedia. 2020a. Dispute tag template. https://en.wikipedia.org/wiki/Template:Disputed.

Wikipedia. 2020b. Npov tag template. https://en.wikipedia.org/wiki/Template:POV.

Wikipedia. 2020c. Wikipedia dispute resolution. https://en.wikipedia.org/wiki/Wikipedia:Dispute_resolution.

Wikipedia. 2020d. Wikipedia personal security. https://en.wikipedia.org/wiki/Wikipedia:Personal_security_practices.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th International Conference on World Wide Web*, pages 1391–1399.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 1480–1489.

Dawei Yin, Zhenzhen Xue, Liangjie Hong, Brian D Davison, April Kontostathis, and Lynne Edwards. 2009. Detection of harassment on web 2.0. *Proceedings of the Content Analysis in the WEB*, 2:1–7.

Justine Zhang, Jonathan P Chang, Cristian Danescu-Niculescu-Mizil, Lucas Dixon, Yiqing Hua, Nithum Tahin, and Dario Taraborelli. 2018. Conversations gone awry: Detecting early signs of conversational failure. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics.*, volume 1.

Justine Zhang, Ravi Kumar, Sujith Ravi, and Cristian Danescu-Niculescu-Mizil. 2016a. Conversational flow in Oxford-style debates. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 136–141, San Diego, California. Association for Computational Linguistics.

Justine Zhang, Ravi Kumar, Sujith Ravi, and Cristian Danescu-Niculescu-Mizil. 2016b. Conversational flow in oxford-style debates. In *Proceedings of NAACL-HLT*, pages 136–141.

A Words list

Words associated with the positive and negative class are shown in Table 5. Coefficients are determined by training a bag-of-words model with logistic regression. The positive class is escalation.

Words	Coefficients
npov	-8.421
information	6.138
did	5.583
wp	4.969
right	-4.819
agree	-4.124
discussion	-4.106
issue	3.864
say	3.771
pov	-3.754
want	3.610
article	-3.398
work	3.362
articles	2.755
edit	2.747
claim	-2.232
case	2.180
people	-2.140
wikipedia	2.132
use	-2.013

Table 5: Words associated with the positive and negative class, using a bag-of-words model to predict escalation.