Stance Classification, Outcome Prediction, and Impact Assessment: NLP Tasks for Studying Group Decision-Making

Elijah Mayfield and Alan W Black
Language Technologies Institute
Carnegie Mellon University
elijah@cmu.edu, awb@cs.cmu.edu

Abstract

In group decision-making, the nuanced process of conflict and resolution that leads to consensus formation is closely tied to the quality of decisions made. Behavioral scientists rarely have rich access to process variables, though, as unstructured discussion transcripts are difficult to analyze. Here, we define ways for NLP researchers to contribute to the study of groups and teams. We introduce three tasks alongside a large new corpus of over 400,000 group debates on Wikipedia. We describe the tasks and their importance, then provide baselines showing that BERT contextualized word embeddings consistently outperform other language representations.

1 Introduction

In the study of groups and teams, measuring discussion quality - plainly, what makes a group debate *good*? - is an open research area. Controlled behavioral studies have shown, for instance, that creativity, diversity, and conflict have major roles to play in the quality of teamwork (Caruso and Williams Woolley, 2008). But the value of diverse discussion and open conflict is complicated, with a long history of positive, negative, and null results, depending on the narrow construct being studied (Jehn et al., 1999). What is clear is that the particulars of how teams are composed and how teammates interact with each other matters a great deal for effective group work (Milliken et al., 2003; Kozlowski and Ilgen, 2006).

In behavioral science, questions are often explored through structured equation modeling and multivariate regressions, allowing behavior scientists sophisticated control over exogenous (fixed, external) variables, like demographics and task conditions, as well as *process* variables that describe observable behaviors in the groups being studied (Cheung and Lau, 2008). Reducing team dynamics from text transcripts to quantitative process variables is computationally complex; in practice, text data is often ignored in favor of proxies like count statistics or, more frequently, participant survey responses (Beal et al., 2003).

These proxies are reliable and effective as stand-ins, but put a limit on the types of questions that can be asked. Scientists studying teams may wish to evaluate which voices truly influenced a conversation, gauge the diversity of people or ideas represented in those influential roles, and measure observed conflicts and consensus-building. They may also want to assess whether any particular participant impacted the discussion and use these variables in aggregate to find which processes impact quality. This data is difficult to extract from discussion transcripts.

Of course, large-scale corpus analysis is common in natural language processing, with many efficient representations of the complex underlying meaning of texts. In this work, we use these methods in the domain of group decision-making research, with three tasks for studying groups:

- **Stance[1] classification**, a fine-grained, fully supervised classification task for individual contributions to a discussion.
- **Outcome prediction**, a distantly supervised task requiring far less annotated training data for new domains.
- **Individual impact assessment**, an unsupervised extension of outcome prediction to quantify how individual contributions or users influenced debate.

In the rest of this work, we demonstrate that these tasks are tractable for NLP researchers to-

[1] In other fields, the term “preference” is often used where NLP researchers would say “stance.” Throughout this work, we use these terms mostly interchangeably.

Proceedings of the Third Workshop on Natural Language Processing and Computational Social Science, pages 65–77
Minneapolis, Minnesota, June 6, 2019.

day, especially with modern language representations like BERT (Devlin et al., 2018). This contextual representation is highly accurate in both supervised tasks and produces interpretable results for the unsupervised task, suggesting it is ready for immediate application in social science research. Alongside these results, we also introduce a real-world corpus of over 423,000 debates from Wikipedia, preprocessed and released under an open source license.

2 Background

2.1 Prior Work on Groups

Group discussion data is commonly used in NLP research. Datasets include the multiparty in-person group work of the AMI meeting corpus (McCowan et al., 2005) and the pair task-based dialogues in the MapTask corpora (Anderson et al., 1991; Bard et al., 1996). A range of core tasks have improved based on these corpora, including diarization (Anguera et al., 2012), laughter detection (Petridis and Pantic, 2008), and summarization (Riedhammer et al., 2010). In online contexts, group debates have been analyzed for tasks like argument mining (Mao et al., 2014) and stance classification (Sobhani et al., 2015), among others.

Outside of NLP venues, though, most studies of groups and organizations do not perform sophisticated text mining or analysis. Methods vary; some research focuses on fuzzy logic or economic agent modeling (see Pérez et al. (2018) for a recent systematic review), while others focus on social factors, network analysis, and the interactive aspects of teams (see Levine et al. (1993); Hackman (2011)). Here we do not address open-ended discussions, focusing on task-based debates where multiple people participate, a fixed set of options are available, and there is no gold standard "correct" answer (in social psychology, "Decision-Making Tasks," from McGrath (1984)).

In these tasks, dysfunction leads to poor outcomes. Low-quality group discussion can focus on already-shared knowledge, rather than new problem solving; high-performing groups by contrast have specific characteristics like shared values, mental models, and communication styles, and nuanced patterns of conflict and consensus-building (Stasser and Titus, 1985). But getting at these patterns quantitatively is complex - most social science research instead avoids the question of extracting structure directly from text, instead relying on direct observable variables and survey data (Jehn et al., 1999), or simulation with explicit preferences encoded in modeled agents (Chiclana et al., 2013). In most work on group decision-making (with the notable exception of some collaborative learning settings, see Rosé et al. (2008)), automated discourse analysis is rare.

2.2 Prior Work on Wikipedia

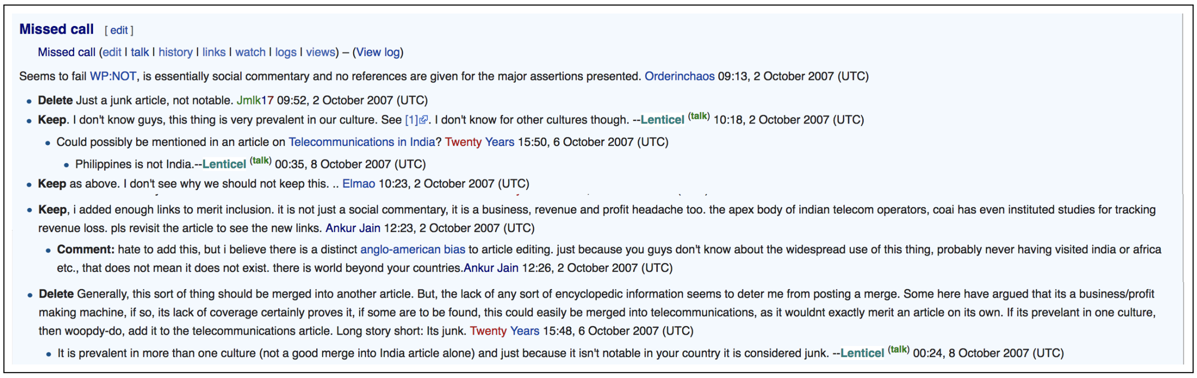
We situate our study in a corpus of Wikipedia data. Ours is far from the first work in this domain, with hundreds of papers published over the last two decades (Mesgari et al., 2015). Large corpora of user discussions on Wikipedia have previously been collected for NLP (Prabhakaran and Rambow, 2016; Hua et al., 2018), though most study discussion in the general case rather than in decision-making contexts. We specifically study *Articles for Deletion* debates (hereafter, *AfD*). In this setting, editors nominate pages for debate that they believe should be removed from the wiki, and other editors debate whether to keep or delete the page. Other reviews of deletion discussions were published during Wikipedia's peak almost a decade ago (Taraborelli and Ciampaglia, 2010; Lam et al., 2010; Geiger and Ford, 2011); since then, only a handful of studies have evaluated this domain, mostly in the context of argument structure mining (Schneider et al., 2013). Since the most recent comprehensive study of *AfD* (Lam et al., 2011), available data has nearly tripled.

The challenges of maintaining good discussion quality are directly applicable to Wikipedia. The health of production communities online requires good working conditions for users, who are volunteers (Halfaker et al., 2011); however, many attempts at improving experiences actually *decrease* productivity and retention (Schneider et al., 2014). The *AfD* process is not amenable to automation or algorithmic decision-making[2]; instead, it is the process itself and the quality of interactions that must be prioritized and improved over time. Better metrics for teamwork could therefore have an immediate effect on the site's policy and practice.

3 Context and Data

When a page is nominated to *AfD*, a group decision-making task begins. Any user (including unregistered users, provided they sign their

[2] Halfaker, personal communication.



Missed call [edit]

Missed call (edit | talk | history | links | watch | logs | views) – (View log)

Seems to fail WP:NOT, is essentially social commentary and no references are given for the major assertions presented. Orderinchaos 09:13, 2 October 2007 (UTC)

- **Delete** Just a junk article, not notable. Jmlk17 09:52, 2 October 2007 (UTC)
- **Keep**. I don't know guys, this thing is very prevalent in our culture. See [1]. I don't know for other cultures though. --Lenticel (talk) 10:18, 2 October 2007 (UTC)
  - Could possibly be mentioned in an article on Telecommunications in India? Twenty Years 15:50, 6 October 2007 (UTC)
    - Philippines is not India.--Lenticel (talk) 00:35, 8 October 2007 (UTC)
- **Keep** as above. I don't see why we should not keep this. .. Elmao 10:23, 2 October 2007 (UTC)
- **Keep**, i added enough links to merit inclusion. it is not just a social commentary, it is a business, revenue and profit headache too. the apex body of indian telecom operators, coai has even instituted studies for tracking revenue loss. pls revisit the article to see the new links. Ankur Jain 12:23, 2 October 2007 (UTC)
  - **Comment:** hate to add this, but i believe there is a distinct anglo-american bias to article editing. just because you guys don't know about the widespread use of this thing, probably never having visited india or africa etc., that does not mean it does not exist. there is world beyond your countries.Ankur Jain 12:26, 2 October 2007 (UTC)
- **Delete** Generally, this sort of thing should be merged into another article. But, the lack of any sort of encyclopedic information seems to deter me from posting a merge. Some here have argued that its a business/profit making machine, if so, its lack of coverage certainly proves it, if some are to be found, this could easily be merged into telecommunications, as it wouldnt exactly merit an article on its own. If its prevelant in one culture, then woopdy-do, add it to the telecommunications article. Long story short: Its junk. Twenty Years 15:48, 6 October 2007 (UTC)
  - It is prevalent in more than one culture (not a good merge into India article alone) and just because it isn't notable in your country it is considered junk. --Lenticel (talk) 00:24, 8 October 2007 (UTC)

Figure 1: Excerpt from a single AfD discussion, with a nominating statement, five votes, and four comments displayed. Votes labeled in **"bold"** are explicit preferences (or stances), which are masked in our tasks.

post with an IP address) can participate, providing either votes or comments. Votes are public, signed, and timestamped; they contain a labeled stance followed by a rationale for why they believe an article should be kept or removed from the wiki. Non-voting *comments*, either in direct reply to the nomination, a vote, or other comments, contain only rationales and not labeled preferences. The structure of these votes and comments follows the standard "reply tree" model of online forums (Aragón et al., 2017).

After nomination, discussions are held open for at least seven days[3]. Discussions are then closed by an administrator, who determines the discussion outcome. While this is not a popular vote, administrators rarely deviate from group consensus. Administrators may also "re-list" debates to hold another seven days of discussion, or close discussions with a verdict of `No consensus`. When that happens, articles are kept by default.

Figure 1 gives an example of how these dynamics play out in practice for the article *"Missed Call."* The nominating statement cites the *"Wikipedia is not a dictionary"* policy and lack of sources to open the debate:

> **Orderinchaos** (nomination): *Seems to fail WP:NOT, is essentially social commentary and no references are given for the major assertions presented.*

This statement is followed by votes and comments, which also contain rationale texts. User preferences for `Delete` and `Keep` are given in **bold**, with some users voting to remove the page, some to keep, and discussion occurring through followup comments:

> **Jmlk** (voting for `Delete`): *"Just a junk article, not notable."*
>
> **Ankur Jain** (voting for `Keep`): *"I added enough links to merit inclusion"* .
>
> **Lenticel** (voting for `Keep`): *"this thing is very prevalent in our culture. See [1]."* .
>
> **Ankur Jain** (comment): *"just because you guys don't know about the widespread use of this thing, that does not mean it does not exist"*

After a long discussion and a total of eight votes and thirteen comments from ten total participants, the decision was made in favor of `Keep`.

3.1 Notation

For any discussion d, we say that it has length N corresponding to the number of contributions $[c_0, c_1, \ldots c_N]$, which are nominating statements, comments, or votes. Each contribution c_i has a corresponding tuple (u_i, t_i, r_i, l_i) representing extracted user, timestamp, rationale text, and stance label, respectively. In our corpus we provide two possible labeling schemes L, a 2-label case for binary classification (which we use), and a 5-label case for direct comparison with prior work like Lam et al. (2010):

$$L_2 = \{\texttt{Delete}, \texttt{Keep}\}$$

$$L_5 = \{\texttt{Delete}, \texttt{Keep}, \texttt{Merge}, \texttt{Redirect}, \texttt{Other}\}$$

In any discussion, the initial contribution c_0 is the nominating statement, which is assumed to have a preference $l_0 = \texttt{Delete}$. For comments, $l_i = \varnothing$. Table 1 gives distributions of these labels in our corpus; the two primary labels dominate. The largest difference between vote and outcome distributions is from `No consensus` results, which default to `Keep` in practice.

[3] Exceptions to this timeline exist and allow "speedy" resolution of discussions - for instance, libelous pages or plagiarism of copyrighted material.

Label	5-Label		2-Label	
	Vote	Final	Vote	Final
Delete	54.9	64.0	62.3	73.0
Keep	28.5	20.5	37.7	27.0
Merge	3.6	3.1		
Redirect	3.8	5.9		
Other	9.3	6.5		

Table 1: Distribution of preference labels in votes and final outcomes in our corpus.

3.2 Corpus and Experimental Details

We evaluate an offline database of all Articles for Deletion discussions[4]. The snapshot contains approximately 19 million pages. Over one third of the pages in the administrative `Wikipedia:` namespace are archives related to *AfD*. We include all data from January 1, 2005 to December 31, 2018. Prior to 2005, traffic was low and decision-making dynamics were erratic, while data from 2019 is (as yet) incomplete. This 14-year window includes over 423,000 discussions.

For all machine learning results, we train a logistic regression classifier implemented in Scikit-Learn (Pedregosa et al., 2011) with L2 regularization and the LIBLINEAR solver (Fan et al., 2008). Experiments represent average results of 10-fold cross-validation. All instances from a particular discussion appear in only one fold; there is never crossover from the same debate between train and test data. We report results on a randomized subset of 5% of the corpus, approximately 20,000 discussions. In preliminary evaluation, a 20x growth in training data increased computational resources beyond what is practical for social scientists, for model accuracy improvements of less than 1%; we exclude full analyses here but provide training splits (for potential future approaches that benefit from larger corpora) in the released data.

For further details on data release and corpus preprocessing, including how free-form preference labels were collapsed, see Appendix A.

3.3 Language Representations

We consider three representations of language. First, we extract standard binary unigram bag-of-words features $\phi_{BoW}(c)$. These were the standard representation of text data for decades and are still in widespread use (Jurafsky and Martin, 2014).

Bag-of-words models struggle in classification tasks for short texts, where sparsity is a significant problem. The most effective recent solution to this has been word *embeddings*, where words are represented not as a single feature but as dense vectors learned from large unsupervised corpora. This allows similar words to have approximately similar representations, and effectively manages sparsity. In our experiments we test a widely-used and effective word embedding model, GloVe (Pennington et al., 2014), set to the maximum of 300 dimensions and represented as $\phi_{GloVe}(c)$.

The newest word embedding models are *contextual*. Rather than encoding a word's semantics as a static high-dimensional vector, these models adjust the representation of words based on the words they appear near at classification time. This approach, combined with extensive pre-training, has led to improvements on numerous tasks. We use the most effective model to date, the $BERT_{BASE}$ model from Devlin et al. (2018) with 768-dimensional embeddings $\phi_{BERT}(c)$. This model was already trained on Wikipedia texts (and other sources), so we perform no fine-tuning[5].

4 Turn-Level Stance Classification

In most other collaborative team decision-making contexts, opinions are expressed but explicit stances are latent. Because of the unique format of Wikipedia discussions, those stances are easily extracted from **'''bolded'''** votes. We use this as a test case for building supervised classifiers which elicit participant stance based on their statements alone. All bolded text is masked from rationales and models must predict what vote is associated with a given rationale.

Similar tasks have been effective in labeling turns in prose text (see Wilson et al. (2005) and other work with their MPQA corpus), open-ended group dialogues (Stolcke et al., 2000; Mu et al., 2012), and in stance classification for more open-ended social media (Sobhani et al., 2015); here we apply the task to contributions in a structured group decision-making context.

Fundamentally this is a test of how closely the Wikipedia domain hews to other decision-making contexts. If rationales are *not* sufficient to predict

[4] From the January 1, 2019 snapshot `dumps.wikimedia.org/enwiki/20180701`

[5] This may mean text from our corpus is included in $BERT_{BASE}$ training data, causing a minuscule exposure to test data in our experimental setup; we do not investigate this question here, but note it as a complicating factor.

Representation	Accuracy	
	%	κ
Majority Class	63.8	0.00
GloVe	76.0	0.45
Bag-of-Words	81.8	0.59
BERT	82.0	0.60

Table 2: Accuracy of stance classification models for individual contributions, based on rationale text alone.

stances accurately, it means one of two things. Either rationales do not carry information about user preferences, and so are not comparable to group decision-making in contexts where those preferences are not explicitly labeled with votes; or the rationales do carry this information, but they are not tractable with current NLP methods. To evaluate this, we define a task to label each vote in each *AfD* discussion:

- **Possible Labels:** $L = \{\texttt{Delete}, \texttt{Keep}\}$
- **Input:** Rationale text r_i from a single vote.
- **Features:** A representation vector $\phi(c_i)$.
- **Output:** A predicted stance $l \in L$.

We exclude the problem of classifying stance in non-voting comments from our analysis, as no gold labels are available for supervised training. Expansion to distant supervision, where user stances from votes are used as gold labels for that user's comments, is a possibility for future work.

4.1 Results

User stances are explicitly given by users in the original corpus and there is no ambiguity; the upper bound for this task is 100% accuracy and $\kappa = 1.0$. Individual votes or comments have short rationales, however, typically only a sentence or a few words. Despite this, n-gram models provide a robust baseline, and while the BERT model outperforms a unigram baseline, the difference is small. Comparing embeddings, the newer contextualized BERT model outperforms GloVe by more than 6% absolute and 10% relative. Overall, we find that this task is tractable, with good accuracy.

5 Discussion-Level Outcome Prediction

The prior task is a useful proof-of-concept that text rationales carry recognizable stance information and can be reliably recognized. With that being said, the task has limitations for practical use in other group decision-making research. Foremost, it requires training data with labeled votes; this is difficult to get in many cases. Moreover, the stances of individual votes in a discussion are too granular for process variables that aim to represent discussion dynamics overall.

A more relevant goal for social scientists is analysis of group discussions where the preferences of individuals are *unlabeled*, even in training data. Next, we aim to predict the consensus preference of a group, after discussion. This task measures whether language representations can model the many turns in a discussion and mimic the behavior of administrators. To do this, we give as input the rationale texts of nominations, votes, and comments throughout a discussion, and treat the label from administrative closure of a debate as the *only* supervised label of group consensus.

- **Possible Labels**: $L = \{\texttt{Delete}, \texttt{Keep}\}$
- **Input**: Discussion d, with nomination c_0, followed by votes and comments $c_1 \ldots c_N$. Each contribution c_i consists of:
  - User ID u_i.
  - Timestamp t_i.
  - Rationale text r_i.
  - Stance label $l_i \in L$, or for comments, $l = \varnothing$. In experiments other than our gold-label comparison, l_i is masked.
- **Features:** A representation vector $\phi(d)$.
- **Output**: An outcome label $l \in L$.

For our embedding representations, we again extract features ϕ_{GloVe} and ϕ_{BERT}, but in this case there is a need to combine vectors from multiple contributions $[c_0, c_1, \ldots c_N]$ into a single vector for discussion d. To do so, we encode each contribution's rationale r_i separately (again removing all occurrences of **"'bolded"'** text to mask votes). We then average each contribution's vector, normalized for length:

$$\phi(d) = \frac{\sum_{i=0}^{N} \frac{\phi(c_i)}{\ln(len(r_i))}}{N}$$

Unlike in the first task, outcome prediction is distantly supervised and the task is sometimes undecidable; as discussed previously, administrators occasionally close conversations with results of

Representation	Final		Real-Time	
	%	κ	%	κ
Majority Class	74.0	0.00	62.1	0.00
GloVe	81.7	0.49	69.1	0.31
Bag-of-Words	84.2	0.58	72.4	0.39
BERT	85.8	0.62	73.4	0.41
Gold Inputs	93.5	0.83	79.7	0.55

Table 3: Accuracy of outcome prediction models, for full discussions and in real-time predictions.

`No consensus`. To evaluate an upper bound on model accuracy with masked preferences, we include a gold feature vector $\phi^*(d)$ where gold-standard user preference labels *are* made available for modeling. Specifically, for each possible $l \in L$, this vector includes the raw count and percent of votes that label received. While Wikipedia is not a direct democracy, administrators rarely deviate from consensus; this represents a good approximation of an upper bound on meaning representation from rationales alone.

As in the first task, we compare binary bag-of-words, GloVe, and BERT representations. We evaluate these models in two scenarios. First, we consider the case where we only predict outcomes after a full discussion has elapsed (**Final**). Second, we consider a just-in-time classifier that predicts the outcome separately after *each* contribution to the discussion (**Real-Time**). While training data includes only final discussions, N separate instances are generated for testing. As such, long discussions have more influence on reported accuracy. By extension, discussions resulting in `Keep` are also over-weighted, as they tend to have more contributions.

5.1 Results

Table 3 shows our comparison of models. As expected, the model given access to stances of group members is highly accurate. That model is able to predict outcomes with a Cohen's $\kappa = 0.84$ for full discussions. The BERT model also reaches good levels of agreement, outperforming other language representations by at least 1.6% absolute. In the real-time evaluation, GloVe and bag-of-words models are more competitive, but BERT maintains the highest accuracy. All models (including the gold-standard) see significant performance degradation, suggesting that discussions are *not* foregone conclusions after early contributions. To demonstrate this more clearly, see Figure 2, where in conversations of any length, outcome prediction early in the debate is less reliable, then improves in accuracy steadily over time as more contributions are made visible to the classifier.

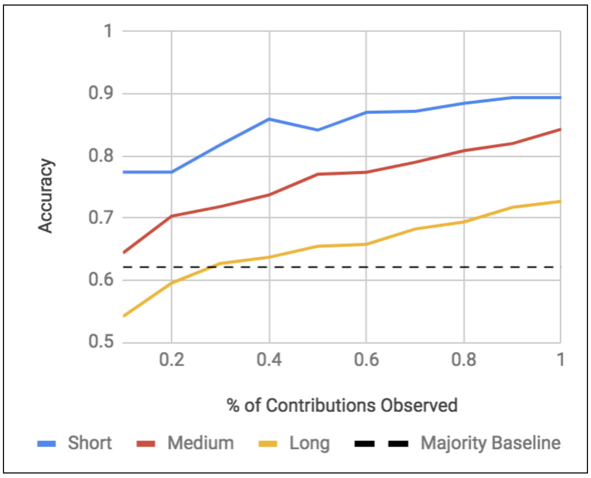


Figure 2: Real-Time BERT model accuracy mid-discussion, split by final debate length: short (5 or fewer), medium (6-10), and long (over 10).

6 Assessing Individual Impact

The prior two tasks were important for understanding how participants use language, and whether preferences of an individual or group are revealed through rationale texts in a discussion. Next, we aim to provide more direct value to behavioral science research, by constructing a metric $Impact(u)$ to identify the primary sources of influence in these discussions.

Our definition of impact hinges on the idea that influential contributions immediately change the likely outcome of a debate. As the basis for this measurement we follow Chouldechova (2017). That work defined "disparate impact" as the difference in expected outcomes, given circumstances that differed by exactly one variable. We borrow this definition, and evaluate impact by varying only time; specifically, we measure the expected outcome of a discussion immediately before and after each contribution is posted[6]. To do so, we use the trained model from our outcome prediction task, in the real-time setting. For a discussion d at a timestamp t, this gives the expected outcome label l - represented as $\mathbb{E}(l|d,t)$, using

[6] For the special case of nominations ($i = 0$), for each possible $P(l)$, for $l \in L$, we instead subtract the baseline probability distribution of all outcomes $l \in L$ as measured from training data.

the model trained on BERT (ϕ_B). Thus:

$$\Delta(l, c_i) = \mathbb{E}(l|d, t_i) - \mathbb{E}(l|d, t_{i-1})$$

Probability movement in one label shifts that label upward, and another simultaneously downward, doubling the cumulative impact of changes; therefore, we sum the change in expected outcomes of all labels and introduce a normalizing factor of ½ to produce an impact value for each contribution ranging from [0,1].

$$Impact(c_i) = \frac{1}{2} \sum_{l \in L} |\Delta(l, c_i)|$$

Finally, we define impact for a user in a conversation as the sum of impacts of their contributions.

$$Impact(u) = \sum_{i=0}^{N} \left\{ \begin{matrix} Impact(c_i) & u_i = u \\ 0 & otherwise \end{matrix} \right\}$$

This measurement of impact, based on probabilities learned from outcome prediction, again does *not* require any explicit labeling on the level of individuals or turns.

6.1 Evaluation

The prior two tasks were supervised, with labeled outcomes that could be measured for performance accuracy. Impact assessment has no specific ground truth to compare against. In this scenario, other NLP research has provided justifications for a mix of quantitative and qualitative evaluations, as well as validation with human annotators and evaluation based on performance improvement in downstream tasks (Louis and Nenkova, 2013; Yang, 2019). We present a mix of qualitative analysis and downstream tasks, while leaving room for future validation studies.

6.1.1 Application: Measuring Volatility

Wikipedia's cultural preference is for open debate and a willingness to voice contrasting views (discussions should be *"not a mere formality, but an integral part of writing the encyclopedia"*[7]). Using raw activity counts cannot measure this, partially because many contributors join late in discussions, mostly to voice agreement for foregone conclusion outcomes, a result of social rewards for editors who participate in more discussions and increase edit counts (Derthick et al., 2011).

To avoid reliance on counts, we define the volatility of a full discussion as the total amount of impact in that discussion, over all contributions:

$$Volatility(d) = \sum_{i=0}^{N} Impact(c_i)$$

We find that our this measure is effective in capturing highly contentious debates. As an illustration, the most volatile debate in our corpus was on the article[8], *"Justin Bieber on Twitter"*. This debate which resulted in nearly 100 votes and many more comments, an article rewrite, a followup deletion review (an appeals process meant to serve an oversight function for *AfD*), an extended external debate on Wikipedia's general purpose discussion board, and the establishment of prevailing policy thereafter for *"[X] on Twitter"* articles.

Other long debates with similarly large numbers of participants were given low volatility scores based on our outcome prediction metric. Upon inspection, these debates end with a string of repetitive votes, like *"**Delete.** As per nom."* or *"**Delete.** As above."* While these debates have high counting statistics, each late vote alters the expected outcome probabilities by well under 0.1%, and sometimes by less than 1×10^{-5}%. This is an intuitive result, accurately reproducing the qualitative findings on behaviors from Derthick et al. (2011). Additionally, this means that more "talkative" users with many contributions do not necessarily make a greater impact, even though $Impact$ is a running sum rather than normalized by contribution count.

6.1.2 Application: Long-term Roles

We can also use outcome prediction to measure the role specific users play over time. By summing influence across all discussions, we find users who have had a disproportionate impact on the *AfD* process over Wikipedia's lifespan. Ranked highly, we find users like **TenPoundHammer** - a user influential enough to spawn an eponymous and well-cited policy essay[9], *"TenPoundHammer's Law."* The most impactful posts typically occur early in debates and are closely tied to policy (linked in double square brackets) and the broader context of Wikipedia's social norms.

[7] en.wikipedia.org/wiki/Wikipedia:Dispute_resolution

[8] bit.ly/2FcSNY7

[9] en.wikipedia.org/wiki/Wikipedia:TenPoundHammer%27s_Law

> (voting for `Delete`) *"Unlikely redirect term, hasn't charted yet, the sources above only confirm that the single "will be" released and tell absolutely nothing else about it."*
>
> (voting for `Keep`) *"I would think anyone who played the NFL for ten seasons is notable... [[WP:Notability (people)]] seems to suggest so."*

We can also evaluate *average* impact rather than cumulative. Here, we find users who are highly active and attentive to new debates, participating early in discussions. Our definition does not encode time explicitly, but in practice early contributions have a larger impact. This is particularly true for `Keep` votes to open debate, which are uniquely influential. Posting in favor of `Keep` immediately after a nomination influences probabilities by nearly three times as much as early `Delete` votes, and more than five times as much as votes that are the tenth or later contribution to a debate.

We can also find prolific users whose roles nevertheless do *not* had an impact on decision-making. User **Captain Raju**, for instance, is a highly active user primarily participating in administrative tasks like vandalism prevention and sorting, rather than voting. Despite frequent activity, their posts have an $Impact$ measure of less than 2% on average. This matches the past finding of "mopping up" roles, which have high importance for the site and highly active users despite relatively low prestige (Burke and Kraut, 2008; Yang et al., 2017). The BERT-powered metric may therefore be useful for role identification.

Overall, our findings show that our $Impact(u)$ rating matches intuitions when given concrete examples, and is able to give interesting insights into group decision-making dynamics longitudinally and in specific circumstances.

7 Discussion

7.1 Opportunities for NLP

Our error analysis shows that on top of support for social sciences, the remaining errors in classification will only be resolved with improved NLP methods. For instance, in stance classification, there are some cases where individual contributions simply lack the content that is necessary to classify them accurately (e.g. *"Per all the above."*). These cases would benefit from a more detailed awareness of threads of conversation (Zhang et al., 2018). Even more often, classification errors occur when users *themselves* express uncertainty:

Final	ϕ	Short	Medium	Long
Delete	BERT	92.9	85.6	74.7
	Gold	97.3	92.9	85.4
		(-4.4)	*(-7.3)*	*(-10.7)*
Keep	BERT	71.9	80.6	75.0
	Gold	91.8	92.2	85.3
		(-19.9)	*(-11.6)*	*(-10.3)*

Table 4: Accuracy of outcome prediction, split by final outcome and total debate length (as in Figure 2).

> (voting for `Delete`) "[. . .] *as I said, I am not really qualified to assess these sources in a deeper way, other than to indicate their existence, and "apparent" reliability under our usual sourcing guidelines."*

Instances like these require not just classification for stance but also for uncertainty (Forbes-Riley and Litman, 2011). Multi-task learning is a particularly fruitful domain for neural methods and the public release of our full corpus should be a resource for development of that field.

In outcome prediction, we find that text models underperform the gold-labels model when predicting an outcome of `Keep`, particularly for short debates. As seen in Table 4, when predicting `Delete` in short discussions, the BERT model is almost always accurate; as conversations grow, `Delete` predictions become less reliable, at just over 75% for debates longer than 10 contributions.

By contrast, when BERT predicts `Keep` it becomes *more* accurate as conversations grow. In short discussions where the final outcome was `Keep`, performance is at its worst, with a gap in accuracy over 22% compared to the gold model. In conjunction with our $Impact$ metric evaluation, this suggests that there is significant opportunity to better identify *persuasive* early `Keep` votes, which are elusive in existing representations.

Further technological advances may also focus on recognizing short discussions that *ought* to be enhanced with additional evidence, either through intelligent routing to potential participants or direct intervention with relevant content. When the outcome prediction expects a `Keep` decision and few users have participated, there is an opportunity for the gap in debate to be filled with decision support aids showcasing the potential of NLP.

7.2 Further Validation of Impact Measures

Our work evaluating impact as a metric, using downstream interpretation tasks as a measure of success, is preliminary. Prior work in the NLP community has developed evaluation metrics hand-in-hand with human input, aiming for high correlation with their judgments (cf. Papineni et al. (2002); Banerjee and Lavie (2005) in machine translation, and Lin (2004) in summarization). This is a natural next step for this work.

Once validated, impact assessment has immediate applications. Distinguishing the impact of individuals will enable deeper process analysis of the impact of diversity on teams (Bear and Williams Woolley, 2011), the interplay between individual participants and the process of resolving conflicts or disputes (Jehn et al., 1999), and the granular habits that lead to effective outcomes. These habits are often process-oriented, small-scale, and not adequately captured by survey or demographic variables (Riedl and Williams Woolley, 2017), opening exciting new dimensions for behavioral science research.

References

Anne H Anderson, Miles Bader, Ellen Gurman Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, Stephen Isard, Jacqueline Kowtko, Jan McAllister, Jim Miller, et al. 1991. The hcrc map task corpus. *Language and speech*, 34(4):351–366.

Xavier Anguera, Simon Bozonnet, Nicholas Evans, Corinne Fredouille, Gerald Friedland, and Oriol Vinyals. 2012. Speaker diarization: A review of recent research. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(2):356–370.

Pablo Aragón, Vicenç Gómez, David García, and Andreas Kaltenbrunner. 2017. Generative models of online discussion threads: state of the art and research challenges. *Journal of Internet Services and Applications*, 8(1):15.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Ellen Gurman Bard, Catherine Sotillo, Anne H Anderson, Henry S Thompson, and Martin M Taylor. 1996. The dciem map task corpus: Spontaneous dialogue under sleep deprivation and drug treatment. *Speech Communication*, 20(1-2):71–84.

Daniel J Beal, Robin R Cohen, Michael J Burke, and Christy L McLendon. 2003. Cohesion and performance in groups: A meta-analytic clarification of construct relations. *Journal of applied psychology*, 88(6):989.

Julia B Bear and Anita Williams Woolley. 2011. The role of gender in team collaboration and performance. *Interdisciplinary science reviews*, 36(2):146–153.

Moira Burke and Robert Kraut. 2008. Mopping up: modeling wikipedia promotion decisions. In *Proceedings of the ACM conference on Computer supported cooperative work*, pages 27–36. ACM.

Heather M Caruso and Anita Williams Woolley. 2008. Harnessing the power of emergent interdependence to promote diverse team collaboration. In *Diversity and groups*, pages 245–266. Emerald Group Publishing Limited.

Gordon W Cheung and Rebecca S Lau. 2008. Testing mediation and suppression effects of latent variables: Bootstrapping with structural equation models. *Organizational research methods*, 11(2):296–325.

Francisco Chiclana, JM Tapia GarcíA, Maria Jose del Moral, and Enrique Herrera-Viedma. 2013. A statistical comparative study of different similarity measures of consensus in group decision making. *Information Sciences*, 221:110–123.

Alexandra Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163.

Katie Derthick, Patrick Tsao, Travis Kriplean, Alan Borning, Mark Zachry, and David W McDonald. 2011. Collaborative sensemaking during admin permission granting in wikipedia. In *International Conference on Online Communities and Social Computing*, pages 100–109. Springer.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. Liblinear: A library for large linear classification. *Journal of machine learning research*, 9(Aug):1871–1874.

Kate Forbes-Riley and Diane Litman. 2011. Benefits and challenges of real-time uncertainty detection and adaptation in a spoken dialogue computer tutor. *Speech Communication*, 53(9-10):1115–1136.

R Stuart Geiger and Heather Ford. 2011. Participation in wikipedia's article deletion processes. In *Proceedings of the 7th International Symposium on Wikis and Open Collaboration*, pages 201–202. ACM.

J Richard Hackman. 2011. *Collaborative intelligence: Using teams to solve hard problems*. Berrett-Koehler Publishers.

Aaron Halfaker, Aniket Kittur, and John Riedl. 2011. Don't bite the newbies: how reverts affect the quantity and quality of wikipedia work. In *Proceedings of the 7th international symposium on wikis and open collaboration*, pages 163–172. ACM.

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. Association for Computational Linguistics.

Karen A Jehn, Gregory B Northcraft, and Margaret A Neale. 1999. Why differences make a difference: A field study of diversity, conflict and performance in workgroups. *Administrative science quarterly*, 44(4):741–763.

Dan Jurafsky and James H Martin. 2014. *Speech and language processing*, volume 3. Pearson London.

Steve WJ Kozlowski and Daniel R Ilgen. 2006. Enhancing the effectiveness of work groups and teams. *Psychological science in the public interest*, 7(3):77–124.

Shyong K Lam, Jawed Karim, and John Riedl. 2010. The effects of group composition on decision quality in a social production community. In *Proceedings of the 16th ACM international conference on Supporting group work*, pages 55–64. ACM.

Shyong K Lam, Anuradha Uduwage, Zhenhua Dong, Shilad Sen, David R Musicant, Loren Terveen, and John Riedl. 2011. Wp: clubhouse?: an exploration of wikipedia's gender imbalance. In *Proceedings of the 7th international symposium on Wikis and open collaboration*, pages 1–10. ACM.

John M Levine, Lauren B Resnick, and E Tory Higgins. 1993. Social foundations of cognition. *Annual review of psychology*, 44(1):585–612.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Proc. of Workshop on Text Summarization Branches Out, Workshop at ACL 2004*.

Annie Louis and Ani Nenkova. 2013. Automatically assessing machine summary content without a gold standard. *Computational Linguistics*, 39(2):267–300.

Fiona Mao, Robert Mercer, and Lu Xiao. 2014. Extracting imperatives from wikipedia article for deletion discussions. In *Proceedings of the First Workshop on Argumentation Mining*, pages 106–107.

Iain McCowan, Jean Carletta, Wessel Kraaij, Simone Ashby, S Bourban, M Flynn, M Guillemot, Thomas Hain, J Kadlec, Vasilis Karaiskos, et al. 2005. The ami meeting corpus. In *Proceedings of the 5th International Conference on Methods and Techniques in Behavioral Research*, volume 88, page 100.

Joseph Edward McGrath. 1984. *Groups: Interaction and performance*, volume 14. Prentice-Hall Englewood Cliffs, NJ.

Mostafa Mesgari, Chitu Okoli, Mohamad Mehdi, Finn Årup Nielsen, and Arto Lanamäki. 2015. the sum of all human knowledge: A systematic review of scholarly research on the content of wikipedia. *Journal of the Association for Information Science and Technology*, 66(2):219–245.

Frances J Milliken, Caroline A Bartel, and Terri R Kurtzberg. 2003. Diversity and creativity in work groups. *Group creativity: Innovation through collaboration*, pages 32–62.

Jin Mu, Karsten Stegmann, Elijah Mayfield, Carolyn Rosé, and Frank Fischer. 2012. The acodea framework: Developing segmentation and classification schemes for fully automatic analysis of online discussions. *International journal of computer-supported collaborative learning*, 7(2):285–305.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830.

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.

Ignacio J Pérez, Francisco Javier Cabrerizo, Sergio Alonso, YC Dong, Francisco Chiclana, and Enrique Herrera-Viedma. 2018. On dynamic consensus processes in group decision making problems. *Information Sciences*, 459:20–35.

Stavros Petridis and Maja Pantic. 2008. Audiovisual discrimination between laughter and speech. In *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5117–5120. IEEE.

Vinodkumar Prabhakaran and Owen Rambow. 2016. A corpus of wikipedia discussions: Over the years, with topic, power and gender labels. In *LREC*.

Korbinian Riedhammer, Benoit Favre, and Dilek Hakkani-Tür. 2010. Long story short–global unsupervised models for keyphrase based meeting summarization. *Speech Communication*, 52(10):801–815.

Christoph Riedl and Anita Williams Woolley. 2017. Teams vs. crowds: A field test of the relative contribution of incentives, member ability, and emergent collaboration to crowd-based problem solving performance. *Academy of Management Discoveries*, 3(4):382–403.

Carolyn Rosé, Yi-Chia Wang, Yue Cui, Jaime Arguello, Karsten Stegmann, Armin Weinberger, and Frank Fischer. 2008. Analyzing collaborative learning processes automatically: Exploiting the advances of computational linguistics in computer-supported collaborative learning. *International journal of computer-supported collaborative learning*, 3(3):237–271.

Jodi Schneider, Bluma S Gelley, and Aaron Halfaker. 2014. Accept, decline, postpone: How newcomer productivity is reduced in english wikipedia by pre-publication review. In *Proceedings of the international symposium on open collaboration*, page 26. ACM.

Jodi Schneider, Krystian Samp, Alexandre Passant, and Stefan Decker. 2013. Arguments about deletion: How experience improves the acceptability of arguments in ad-hoc online task groups. In *Proceedings of the 2013 conference on Computer supported cooperative work*, pages 1069–1080. ACM.

Parinaz Sobhani, Diana Inkpen, and Stan Matwin. 2015. From argumentation mining to stance classification. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 67–77.

Garold Stasser and William Titus. 1985. Pooling of unshared information in group decision making: Biased information sampling during discussion. *Journal of personality and social psychology*, 48(6):1467.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational linguistics*, 26(3):339–373.

Dario Taraborelli and Giovanni Luca Ciampaglia. 2010. Beyond notability. collective deliberation on content inclusion in wikipedia. In *Self-Adaptive and Self-Organizing Systems Workshop (SASOW), 2010 Fourth IEEE International Conference on*, pages 122–125.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*.

Diyi Yang. 2019. *Computational Social Roles*. Ph.D. thesis, Carnegie Mellon University.

Diyi Yang, Aaron Halfaker, Robert Kraut, and Eduard Hovy. 2017. Identifying semantic edit intentions from revisions in wikipedia. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2000–2010.

Justine Zhang, Danescu-Niculescu-Mizil, Christy Sauper, and Sean Taylor. 2018. Characterizing online public discussions through patterns of participant interactions. In *Proceedings of CSCW*.

A Appendix

A.1 Corpus Preprocessing

Compared to the broader internet, Wikipedia is simpler to preprocess due to the rigid formality of the archival process, the MediaWiki markup language, and enforced community standards. For most tasks, we are able to extract names, timestamps, and labels with only regular expressions.

Extracting Timestamps

AfD discussion norms require that all contributions are signed using a standard format, which includes the contributor's username or IP address and a timestamp in UTC format[10]. All lines following the outcome are checked for timestamps in Wikipedia standard format[11]:

```
\d\d:\d\d, \w+ \d+, 20\d\d (UTC)
```

Extracting outcomes

AfD discussions are archived in a specific format with only minor variation, and can be easily extracted for structured representation. We define a discussion as having an *outcome* if its archival page includes a header line with one of three fixed phrases (ignoring whitespace):

```
The result of the debate was [x]
The result was [x]
The result of this discussion was [x]
```

We save the captured string `[x]` as the debate outcome. When these lines are timestamped, we also log the user and timestamp of the outcome.

Extracting nominations, votes, and comments.

If a timestamped contribution appears at the top of the discussion, prior to any votes, it is treated as a *nomination*. These statements have become more common over time: while they occur in only 67% of nominations in 2005, they were rapidly adopted and are present in 98% of nominations since 2008[12].

Following the nominating statement, any timestamped line is captured as either a vote or a comment. We define votes as any timestamped line beginning with a bolded phrase, following Wikipedia convention for contributions:

```
* ```[y]'''
```

Posts beginning with one or more leading asterisks creates a bulleted, threaded discussion. Words or phrases surrounded with three apostrophes creates '''**bolded**''' text. The value of this bolded text `[y]` is captured and stored. If no bolded phrase is present, but the line is still signed and timestamped, that line is treated as a *comment*[13]. Lines with no timestamped signature are discarded.

Several alternative solutions to deletion exist; each maintains the content of the page while deleting the page itself. In the five-label case, `Merge` and `Redirect`, the two most common alternate outcomes, are represented separately in line with prior work; in the two-label case they are merged in with Delete. All other values are grouped together as `Other` in the five-label case[14]; in the two-label case they are merged in with `Keep`. Votes and outcomes of "Close", "Withdraw", and "Cancel" are treated as "Keep" outcomes as the page as well as its content is fully maintained. Copyright violations are treated as a "Delete" outcome, as the content is deleted as a result of the outcome. Any given vote or outcome is represented as a set that can contain zero or more normalized labels. Therefore, the probability of a vote for a particular label is not drawn from a distribution; probabilities of each label in L are disjoint.

Extracting users

For each nomination, outcome, vote, or comment, we log the user whose signature immediately appears before the timestamp, either with a MediaWiki link to their User page or their User Talk page:

```
[[User Talk:[z]
[[User:[z]
```

We extract `[z]` as a username and associate it with the nomination, outcome, vote, or comment where it was captured. When user signatures link to both User and User Talk pages and those usernames differ, the Talk page's username is prioritized.

[10] These signatures are highly formulaic and easy to extract, because they can be automatically generated by MediaWiki's `~~~~` shorthand. When users do not sign contributions, bots add them, along with a citation to the `SIGNATURES` policy.

[11] In regular expressions, `\w` matches any letter and `\d` can match any numeric character. A + suffix captures one or more consecutive characters of that type.

[12] Under present policy, omitting a nominating statement is an acceptable reason for "speedy" dismissal and default "Keep" outcome for an AfD nomination.

[13] Lines beginning with the bolded phrase **"Comment"** are also treated as comments. Lines beginning with **"Note"** are automatically generated, typically for categorizing discussions by topic, and are discarded. Lines with "Relist" bolded are administrative notes to keep the discussion open for longer than the typical seven days, and are also discarded.

[14] "Userfy", "Transwiki", "Move", and "Incubate"

A.2 Reproducibility

The public release of this corpus will include designated fold assignments for reproducible results and future comparisons against baselines on the 5% subset used in this work. We will also include two formats for experimenting with the full corpus: a 10-fold cross-validation split, as well as a single train/validation/test split for use with more resource-intensive classifiers, especially neural methods.

The library that we developed for producing these variables is written in Python and compatible with standard implementation of BERT and a standard JSON format for representing group discussions.