WikiTalkEdit: A Dataset for modeling Editors' behaviors on Wikipedia

Kokil Jaidka[1] **Andrea Ceolin** [2] **Iknoor Singh**[3]
Niyati Chhaya[4] **Lyle H. Ungar**[5]

[1] Department of Communications and New Media, National University of Singapore
[2] Department of Linguistics, University of Pennsylvania
[3] Department of Computer Science, University of Sheffield
[4] Big Data Experience Lab, Adobe Research India Pvt. Ltd.
[5] Department of Computer & Information Science, University of Pennsylvania
jaidka@nus.edu.sg

Abstract

This study introduces and analyzes **WikiTalkEdit**, a dataset of conversations and edit histories from Wikipedia, for research in online cooperation and conversation modeling. The dataset comprises dialog triplets from the Wikipedia Talk pages, and editing actions on the corresponding articles being discussed. The exchanges occur between two turn-taking individuals and span all of Wikipedia. We show how the data supports the classic understanding of style matching, where positive emotion and the use of first-person pronouns predict a positive emotional change in a Wikipedia contributor. However, they do not predict editorial behavior. On the other hand, feedback invoking evidentiality and criticism, and references to Wikipedia's community norms, is more likely to persuade the contributor to perform edits but is less likely to lead to a positive emotion. We developed baseline classifiers trained on pretrained RoBERTa features that can predict editorial change with an F_1 score of .54, as compared to an F_1 score of .66 for predicting emotional change. A diagnostic analysis of persisting errors indicates that deep learning models often fail when political and religious topics are being discussed. We conclude with possible applications and recommendations for future work. The dataset is publicly available for the research community at https://github.com/kj2013/WikiTalkEdit/.

1 Introduction

Dialogue is a language game of influence, action, and reaction that progresses in a turn-taking manner. Persuasion occurs through dialogue when a listener favorably evaluates the authority, claims, and evidentiality through the cues and arguments made by the speaker (Krippendorff, 1993; Schulte, 1980; Durik et al., 2008).

Discussions on Wikipedia Talk pages can be useful for determining strategies that lead to an improvement of the artice discussed, and for examining if they also lead to an amicable dialogic exchange. Previous work (Yang et al., 2016a,b, 2017) has explored the role of editors and the types of edits made on Wikipedia, but have not related them to the ongoing conversation on the Wikipedia Talk pages.

We introduce the **WikiTalkEdit dataset**, a novel dataset for research in online collaboration. The dataset is a subset of the Wikipedia Talk Corpus available as of May 2018[1]. It contains 12,882 dialogue triples with labels about editors' subsequent editorial (editing) behavior, and 19,632 triplets with labels corresponding to editors' emotion as manifested in their replies. Table 1 has examples from the dataset.[2]

This new dataset enables various language and behavior modeling tasks. In general, the dataset is important for understanding linguistic coordination, online cooperation, style matching, and teamwork in online contexts. More specifically, it offers linguistic insights about the norms on Wikipedia, such as (i) the feedback which is associated with a positive emotion vs a positive editing action, (ii) identifying and characterizing successful editorial coordination (Lerner and Lomi, 2019), (iii) generating constructive suggestions based on a given Wikipedia edit, and (iv) identifying and resolving disagreements on Wikipedia before they go awry (Zhang et al., 2018). In this study, we examine the first research problem. That is, we demonstrate how the dataset is helpful to compare and contrast the linguistic strategies that evoke favorable dialogic responses from those evoking behavioral compliance.

[1] https://figshare.com/articles/Wikipedia_Talk_Corpus/4264973

[2] Code to replicate the data collection is available at https://github.com/kj2013/WikiTalkEdit/.

2 Related Work

Conversational quality is largely the focus of a body of work modeling the formal (Pavlick and Tetreault, 2016), polite (Niculae et al., 2015) and toxic (Zhang et al., 2018) features of comments on Wikipedia, Reddit, and other online public forums. The labels in such a task are often subjective as they depend mostly on annotated or crowdsourced labels. On the other hand, gauging the impact of a conversation in terms of a reader's subsequent behavior is a rather different problem. A few studies have modeled the language of arguments to predict their upvotes (Wei et al., 2016a,b; Habernal and Gurevych, 2016; Tan et al., 2016). The best result reported by Habernal and Gurevych (2016) was an F_1 score of .35 for the task of predicting which of two arguments was better, using SVMs and bi-directional LSTMs. The study by Tan et al.(2016) reported an accuracy of 60% for predicting which argument was most likely to change the original poster's (OP's) point of view. Althoff et al. (2014) report an AUC of .67 on predicting the success on $\sim$5700 requests. Studies predicting users' stance (Lin and Utz, 2015; Sridhar et al., 2015) have done better, but do not usually factor in the feedback from a turn-taking partner, during a dialogic exchange. Furthermore, to the best of our knowledge, we did not find an equivalent study to measure the actual subsequent behavior of a conversation partner after a dialogic exchange on social media platforms, forums, or Wikipedia.

In recent years, computational linguistics has developed computational models of dialogic text that predict the emotional responses associated with any utterance. The findings suggest that interacting speakers generally reinforce each others' point of view (Kramer et al., 2014; Rimé, 2007), use emotions to signal agreement, and mirror each other's textual cues (Niculae et al., 2015). On the other hand, predicting behavioral responses is potentially a more challenging task for text modeling and prediction, and it is also less explored in the literature.

The existing research on online turn-taking behavior has focused on modeling emotional reactions, with little interest in predicting actual behavioral change. This research is discussed in more detail in the Supplementary Materials [3]. For now, we contextualize the contributions of this dataset by demonstrating how it is applicable to address the following gaps in the scholarship:

- How well do language models trained on editorial feedback predict subsequent emotional and editorial change?
- What are the linguistic features of editorial feedback which predict emotional change in the person that initiates the discussion (henceforth, OP, original poster)?
- What are the linguistic features of editorial feedback which predict subsequent editorial behavior by the OP?

First, we report the predictive performance on predicting emotional and editorial behavior change from the linguistic features of the comments, using regression baselines and state-of-the-art deep learning models. Performance is evaluated as an F_1 score of predicted labels against the ground truth labels as implemented in *scikitlearn*. Then, we compare the linguistic features associated with emotional change with those associated with subsequent edits. Finally, we offer a diagnostic analysis of the prediction errors observed.

Page	**Talk triplet**
Frontiersmen Camping Fellowship	OP. The title of this article should be "Frontiersmen Camping Fellowship". That is the title on my pins, handbooks and in all literature and on the website. [...] E. *I moved* the article to Frontiersmen Camping Fellowship accordingly. **OP. Thanks!**
David Livingstone	OP. Wasn't Lake Victoria the source of the Nile? Why did I find Lake Victoria, and then a few years later start looking for the source of the Nile? [...] E. As *I* recall, he (and numerous others) *thought/hoped* that Lake Victoria *was* just yet another intermediate source. They *were looking* for the source of Lake Victoria (and hence the Nile), or else to discover that Lake Victoria *flowed* into something else, not the Nile [...] **OP. Interesting, thanks!**
Unparished area	OP. how does this look as a format? i'm pondering recasting it in terms of ceremonial counties of England once i'm done (but then that does have the stockton problem) perhaps better left as-is? [...] E. I think it *makes* sense to use the 1974 counties as that is where they were when they became unparished . *You* could put a note under those ones that have changed counties [...]. *You* could just list the districts alphabetically, of course, and leave the counties out of it. **OP. But the districts have renamed and merged lots - i suppose using the current districts would make sense. Once I'm all done, I think maybe...**
Electro Tone Corporation	OP. This is a page I am beginning because of the troubles I had in obtaining information on an Electro Tone Duet Sixteen for my Hammond organ. [...] E. I've removed the speedy tag for now, to see how the article *develops*. It is important for *you* to *cite* your sources (see WP:Verifiability), and also *note* there *is* a notability threshold (see WP:CORP for companies), so the article may still be deleted if notability cannot be asserted. **OP. The majority of information for this article is my own research, with my photos displayed for verifiability. There is no other source in the world that I am aware of that has this information. [...]**

Table 1: Example talk triplets from the dataset.

3 The WikiTalkEdit dataset

In this dataset, we describe how we collected our data from the Wikipedia Talk dataset and formulated a task around emotional and behavioral actions of an article's editors, who are taking turns in a conversation.

[3] Available at https://github.com/kj2013/WikiTalkEdit/

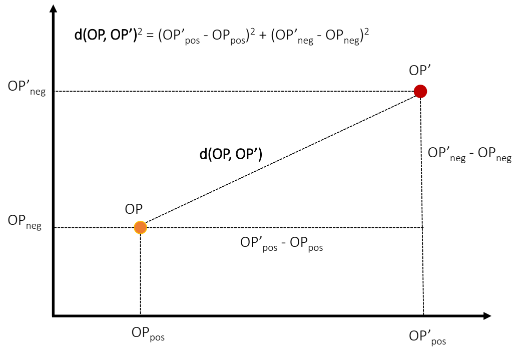

Figure 1: Calculation of OP's emotional change as the signed two-dimensional Euclidean distance between OP and OP'.

	Triplets	0s	1s	Pages	Users
Sentiment change	19,632	7,286	12,346	19,299	12,531
Editorial change	12,882	6,896	5,986	12,731	8,506

Table 2: Dataset statistics.

3.1 Data generation process

After contributing to a Wikipedia article, the OP usually updates the Talk page with a summary of the edit. At this point, the OP may get zero or more responses, and they may respond to all, some, or none of them. To study the effect of editorial feedback, we defined a complete interaction between an OP and another Editor as a dialog triplet of the form $OP \rightarrow Editor \rightarrow OP'$.

Our dependent variables are the OP's reaction to an Editor's comment in terms of the 'emotional change' in their language and their 'editorial change' in terms of subsequent edits to the Wikipedia article.

First we downloaded the entire Wikipedia Talk Corpus available as of May 2018[4] and extracted 128,231 dialogue triplets. Next, we used the Wikimedia API to download the edits corresponding to each of the OP's comments in our dataset of triplets. In the following paragraphs, we further describe how we operationalized the labels for the dataset.

Emotional change: The emotional change label is the **signed** Euclidean distance between the positive and negative emotions of OP' and OP (see Figure 1). The positive and negative emotion measurements are calculated using the emotion dictionaries from the Linguistic Inquiry and Word Count (LIWC) (Pennebaker et al., 2007). The assigned labels were manually examined by the authors for face validity.

A change over one standard deviation above the mean is coded as '1' and is a positive emotional change. A change under one standard deviation below the mean is coded as a '0' and is a negative emotional change. All other values are marked "null" as there is no evident change in emotion.

Editorial change: The edits, if any, performed by the OP to the article in the week following the Editor's feedback, are operationalized as a binary value ('1'=edit, '0'=no edit).

4 Dataset Analysis

In the following sections, we analyze what types of linguistic feedback from the Editor is effective at creating a positive emotional change or an editorial action by the OP.

In preliminary data explorations, we found no correlation between the Editor's politeness or status, and emotional or editorial change. We observed that the Editor's comments that are associated with positive comments (Mean = 273 characters, Mean Jaccard coefficient, JC = .16) are significantly shorter and have less overlap (content interplay) with the OP's comment than those associated with negative comments (Mean = 417 characters, Mean JC = .18). There was no substantial difference for editorial changes.

5 Predicting the response to Editor's Feedback

We examine the different linguistic features and discourse markers in predicting emotional and editorial change in the WikiTalkEdit dataset. Our independent variables comprise the linguistic features of the Editor's feedback, and the dependent variables are the OP's change in emotional and editorial behavior after receiving the feedback.

We used logistic regression and many deep learning implementations from the *pytorch-pretrained-bert* package to predict both the emotional and editorial change of the user.

5.1 Feature extraction

We represented the Editor's feedback as a normalized frequency distribution of the following feature sets:

– **General lexical features (500 features and 50 LDA topics):** The most frequent unigrams, and 50 topics modeled using Latent Dirichlet Allocation in Python's MALLET package, with α=5.

– **Stylistic features (73 features):** These include cognitive features (discrepancy, past tense, present

[4] https://figshare.com/articles/Wikipedia_Talk_Corpus/4264973

tense, work) and emotional features (positive emotion, negative emotion, reward) from LIWC (Pennebaker et al., 2007). A politeness index was generated using Stanford's Politeness API (Danescu-Niculescu-Mizil et al., 2013).

–**Syntactic features (4 features)**: The Stanford parser was used to generate dependency parses. Dependency parses were used to identify and categorize all the adjectival modifiers that occurred at least ten times in the data. We distinguished the first-, second-, and third-person pronouns. Finally, we created part-of-speech n-grams based on the dependency trees.

– **Social features (2 features):** We measured *content interplay*, the Jaccard coefficient of similarity between the unigrams of the Editor's feedback and the OP's first comment. The Editor's *status* may pressure the OP to conform by performing edits; therefore, we quantified the Editor's experience in terms of their number of contributions to Wikipedia.

5.2 Deep learning models

We also experimented with a variety of deep learning baselines:

- CNN: The CNN framework (Kim, 2014) involves applying convolutional filters followed by max-over-time pooling to the word vectors for a post.
- RCNN: The RCNN framework (Lai et al., 2015), recurrent convolutional layers followed by max-pooling. A fully connected layer then follows it with a softmax for output.
- biLSTM: The word embeddings for all words in a post are fed to bidirectional LSTM, followed by a softmax layer for output (Yang et al., 2016c).
- biLSTM-Attention: For each sentence, convolutional and max-over-time pooling layers are applied on the embeddings of its words. The resultant sentence representations are put through bi-LSTM with the attention mechanism (Yang et al., 2016c).
- NeuralMT: Embeddings are fed into a bidirectional-GRU followed by a decoder with the attention mechanism (Bahdanau et al., 2015).
- FastText: Word representations are averaged into a sentence representation, which is, in turn, fed to a linear classifier (Joulin et al., 2017). A softmax function is used to compute the probability distribution over the predefined classes, and a cross-entropy loss is used for tuning. Hierarchical softmax is used to speed up the training process.
- Transformer: The architecture implemented was based on recent previous work (Vaswani et al., 2017).
- OpenAI GPT: The Generative Pretrained Transformer implementation (Radford, 2018) with the original hyperparameter settings.
- BERT and RoBERTa: The pre-trained BERT model (Devlin et al., 2018) and the Robustly optimized BERT model (RoBERTa) (Liu et al., 2019), where BERT is retrained with more data and an improved methodology. Models were fine-tuned using the simple transformers library.
- XLNET: Finally, we evaluate the performance of XLNet (Yang et al., 2019), which combines bidirectional learning with the state-of-the-art autoregressive model such as Transformer-XL.

In the case of CNN and BiLSTM based models, we used the referral hyper-parameters from the original implementation for all models[5]. For Neural MT, FastText, and Transformer based models, implementations by original authors are used. All the models were evaluated using 5-fold cross-validation with a split ratio of 80:20 for train and test set, respectively. In fine-tuning the RoBERTa model on editorial change, the model parameters included a learning rate of 9e^{-6}, 3 epochs, and train batch size of 8. For emotional change, model parameters include a learning rate of 1e^{-5}, 3 epochs, and a train batch size of 8. The maximum input sequence length was 128, which included 91% of all the inputs. The time taken was 6-8 minutes/epoch on Tesla k80, running on a Google Colab implementation. Five hyperparameter search trials were conducted with cross-validation. A manual tuning strategy was followed to identify the setting with the best performance.

6 Results

We now examine the test-set performance of these models trained on a subset of the WikiTalkEdit dataset. The dataset for emotion analysis comprises the 15% of overall dataset where editorial feedback yielded a substantial positive or negative change in

[5] https://tinyurl.com/brightmart

the emotion vector (i.e., the emotional change was above or below one standard deviation from the mean). Similarly, the dataset for editorial actions (edits performed) comprises the 10% of the conversations that started within 24 hours since an OP edited the page. A pairwise correlation found no relationship between emotional and editorial change (ρ= .01, p>.1). The dataset statistics are provided in Table 2.

6.1 Predictive performance

Baseline logistic regression models: Table 3 shows that among the baselines that use bags-of-words and features, emotional change is more straightforward to predict than behavioral change, and style provides marginally better predictive performance than content. The best performance was obtained using POS n-grams, with an F_1 score of .57 for predicting emotional change, and of .51 for predicting behavioral change. Unexpectedly, social features were not good predictors of emotional change.

Deep learning models: In comparison to the logistic regression baselines, the deep learning models in Table 4 offer a remarkable predictive advantage, especially for emotional change. The best performing deep learning classifier is trained on pre-trained RoBERTa features and reports an F_1 score of .66 for emotional change and .54 for editorial change.

	Sentiment change	Editorial change
Features	F_1 score	
General lexical features		
Unigrams	.54	**.51**
Topics	.56	**.51**
Stylistic features		
LIWC + Politeness	.55	.48
Syntactic features		
Dependency Parses	.51	.46
Adjectival modifiers	.46	.42
Personal pronouns	.39	**.51**
POS ngrams	**.57**	**.51**
Social features		
Content Interplay + Status	.39	.35

Table 3: Performance of Logistic Regression classifier on different features.

	Sentiment change	**Editorial change**
Model	**F_1 score**	
CNN (Kim, 2014)	.45	.38
biLSTM	.57	.38
biLSTM-Attention (Yang et al., 2016c)	.59	.43
Neural MT (Bahdanau et al., 2015)	.59	.47
RCNN (Lai et al., 2015)	.61	.36
FastText (Joulin et al., 2017)	.65	.51
Transformer (Vaswani et al., 2017)	.48	.50
OpenAI GPT (Radford, 2018)	.64	.50
BERT (Devlin et al., 2018)	.65	.52
RoBERTa (Liu et al., 2019)	**.66**	**.54**
XLNet (Yang et al., 2019)	.65	.53

Table 4: Predictive performance with deep learning classifiers.

6.2 Error analysis

We observed instances of misclassification from the best logistic regression classifier and the XLNet model (Yang et al., 2019). We have diagnosed the likely sources of errors in this section.

6.2.1 False positives in emotional change prediction

We randomly selected an assortment of false positives predicted by a logistic regression classifier and by XLNet and have provided them in Table 5[6]. First, we find that since the logistic regression methods rely heavily on stylistic features, the errors we identified seemed to occur when the style does not match the intended meaning:

- **Feedback about notability and relevance**: In the first example in Table 5, we see that despite the polite feedback, the conversation was not resolved positively and resulted in negative responses.
- **Reverted edits**: Similarly, in conversations where the OP contest their reverted edits, the dialogue appears to regularly derail into further negative replies despite the civility of the Editor's feedback.

The XLNet model did not repeat these particular errors. Its errors, on the other hand, appear to be driven by fact-checks and questions:

- **Fact-checks**: In contradicting the OP with facts and personal opinions, a disagreement is sometimes implied but not obvious. The model predicts a positive emotional change, but the OP responds to the implication with a negative reaction.
- **Counter-questions**: When Editors asked questions of the OP, it appears likely that the OP would turn defensive, even if the response included facts.

6.2.2 False positives in editorial change prediction

Table 6 shows the false positives in predicting editorial change. Starting with the errors from models trained on stylistic features, we observed that in general, the errors centered on:

- **Controversial topics:** The errors arising from logistic classifiers reflect ideological disagreements, often involving hot-button topics

[6] More examples of errors are provided in the Supplementary Materials

such as race and ethnicity. The OP is not likely to change their mind despite what might be a well-reasoned argument from the Editor.

- **Reverted edits:** Dialog around why edits were reverted, or content was removed are usually requests for greater clarity for documentation purposes, and are rarely followed up with edits to the page.

False positives in predicting editorial change by XLNet also appear to arise when feedback is nuanced. Aside from feedback that implicitly discourages further editing, similar to what was observed in Table 5, we also observed other types of feedback that leads to errors by the XLNet model:

- **Opinions:** Editorial feedback that uses opinions rather than facts to persuade the OP appears to lead to an edit rarely, and this was a common error observed among the predicted labels.
- **Mixed feedback:** The models also appear to get confused when the feedback included content from the page as a quote, and included suggestions but made no direct requests.

7 Linguistic insights

Based on the results in Table 3, in this section, we examine the stylistic, lexical, and topical features which best predict emotional and behavioral change. These findings offer us a way to examine whether emotional and editorial change are indeed different, and to compare the results against previous studies which have examined these problems in some capacity.

7.1 Stylistic insights

Comparing the most predictive stylistic and content features suggests that emotional and editorial change have different predictor. Table 7 summarizes the most significant predictors of emotional change based on an ordinary least squares regression analysis. Positive feedback through words related to rewards and positive emotions typically predict a positive emotional change, besides the use of *stance* words (the first person pronoun, *I*) and reference to past experiences (past tense). This finding is in line with the literature (Zhang et al., 2018; Althoff et al., 2014). Conversely, excessive use of adjectival modifiers (e.g., comparative words or words used to emphasize quantity or impact) is associated with a negative emotional change.

The insights look very different for editorial change (Table 8). Second person pronouns and present tense, both of which occur in directed speech, are associated with editorial changes, in sharp contrast with the features that emerged in the analysis of emotional change. Aligned with this, the use of words related to criticism (discrepancy) and work is also among the significant predictors of editorial change. Among the parts of speech, comments about the content (NN, NNP) appear to reduce the likelihood of an editorial change. Except for superlative modifiers, style seems not to be relevant in this case.

These results support previous studies in showing that emotion and politeness do not always signal editorial change (Hullett, 2005; Althoff et al., 2014), as it is true for stylistic markers (Durik et al., 2008), while direct requests (Burke et al., 2007), assertiveness, evidentiality (Chambliss and Garner, 1996) and other content-based features usually perform better. No feature appeared to correlate with both emotional and editorial behavior. Further lexical insights are provided in the Supplementary Materials.[7]

8 Insights from topics

We conducted a Benjamini Hochberg (BH)-corrected Pearson correlation of the topic features of comments by the Editor. We visualize it as a language confusion matrix introduced in recent work (Jaidka et al., 2020) to compare the topics predictive of emotional vs. editorial change of the OP. The word clouds in Figure 2 show the correlation of LDA topics with emotional change on the X-axis, and the correlation with editorial change on the Y-axis. The grey bands depict zones where the topics do not have a statistically significant correlation with either emotional or editorial change. We have distinguished the themes related to content (e.g., *medicine, religion, and ethnicity*) by coloring them in red. The topics in black are related to Wikipedia's content guidelines (i.e., mentions of *NPOV, sources, cite, information*).[8] These themes involve the neutrality (*neutral point of view, NPOV*), general importance (*notability*), and verifiability (*sources, evidence*) of information.

[7] We further tested the effect of only positive or only negative features; we found that positive emotion is a better predictor of emotional change (F1=.42 vs. F1=.45) but not of editorial change (F1=.48 for both positive and negative features).

[8] See https://en.wikipedia.org/wiki/Wikipedia:Five_pillars

Sentiment change - false positives			
LIWC-based classifier		**XLNet**	
Page	**Talk triplet**	**Page**	**Talk triplet**
Kiev	OP: I am so confused! What is the policy regarding names of towns?[...] Editor: Danny, there's no real policy. If there is a name which is widely used for one very **famous** city (e.g. Paris, Rome, Athens), then the article with that name should be about that city. [...] OP': Actually, I have a problem with this. Yes, Paris, Rome, and Athens immediately bring to mind great European cities, but why should we play favorites?[...]	Jeff Kennet	OP: ==Massive POV/CRUFT should be deleted== I recommend restoring my edits. This article is filled with POV and cruft. I did not think they would require community consensus but it is what it is. Also, why no mention of the psycho who |ssaulted him (Fergal Downey)?? Editor: When I looked at your edits **I was just overwhelmed** by the scale of them. Obviously you've [...] **What do you want to remove?** What do you want to add? [...] OP': I guess sometimes I am too bold.
Moors	OP: Several days have gone by and [user] has still failed to give me a logical reason for his reversion. I changed the article back to my previous edit and made some [...] Editor: In other words, you have no consensus for the changes. I'm **not surprised** hasn't replied, since you've got further and further from any actual comment on the content of the article. I'm not even sure [...] OP': Don't *revert due solely to no consensus* any of Wikipedia's basic editing policies? This is the second time I have been reverted without a specific reason why.[...]	Religion in Swaziland	OP: ==Improving the article== I have made some edits to improve the article. Editor: **Can you please explain why** you changed CIA statistics? OP': ==Reverted page== I disagree with the edits by , I therefore reverted it to my last edit on July 15.

Table 5: Error diagnostics for predicting sentiment change. In this case, our models predicted a positive change ('1'), but the sentiment actually turned negative.

Editorial change - false positives			
LIWC-based classifier		**XLNet**	
Page	**Talk triplet**	**Page**	**Talk triplet**
Arabic	OP: The redirect to the disambiguation page is intended to assist the reader in locating []...] Editor :Wikipedia is not a dictionary. Adjectives are not supposed to be disamiguated in this way unless they are truly ambiguous. The noun has very clear presedence over the rather nonstandard adjective. See talk:Arab [...] OP': I think you are doing a disservice to readers. I am not going to get in a revert war with you, *but will ask for input from others.*	Mario Kart	OP: == Consoles? == Consoles are non-portable video game systems, handhelds are portable ones. Should we change the uses of "Console Games" to "Video Games" due to Mario Kart games on both types? [...] Editor: Handhelds are really just a type of video game console. (Thus the name of the article we have on it, handheld game console.) Using just "video game" create would instead create a new problem, because arcade game | the main thing we're using to differentiate from, are also a kind of video games. OP': Ah, okay, I thought there was a big difference. Thanks for clearing this up! [...]
Oona King	OP :OK, I admit that my own recent edit on this point was not the [...] Editor: *The woman did not appear to be black at all (she looks slightly tanned, almost mediterranean) :Why do you think that being Black is about how a person looks? [...] OP': *There is some merit in your argument, but* race is a matter of genetics, as I see you say yourself. I think we would be a lot better off without racial [...]	Knights Templar legends	OP: == Bannockburn == I've replaced a sentence on the Victorian origin of the myth with a short summary of Cooper's research on Burnes.[...] Editor: Thanks. What really concerned me was [...] We need to cite/attribute this. It's pretty rare to use blogs, see WP:RS and WP:VERIFY. [...] OP': Concern shared. Thanks, we're using the same hymnsheet.

Table 6: Error diagnostics for Editorial change. In this case, our models predicted edits ('1'), but no edit was actually made.

Finally, the blue topics are meta-commentary centered around the elements in a Wikipedia article (mentions of *edit, page, title, section*).

Our analysis of the WikiTalkEdit dataset suggests that mentions of Wikipedia's guidelines are associated with a positive editorial change, but a negative emotional change. Suggestions based on evidence are associated with both, a positive editorial *and* a positive emotional change. First, we look at the spread of the content-themed topics around the figure. Some of the topics related to religion (*god, church, christian*) and ethnicity (*israel, people, jewish, indian*) are associated with a negative emotional change (-.06 < r < -.02, p < .05). Content topics related to medical research and health inspire a negative emotional change but a positive editorial change (r = .05, p<.05).

Next, we consider the meta-commentary about page structure (*page, title, move* and *review, section, add*). We observe that these are associated with positive emotional changes (.06 < r < .10, p < .05), possibly because they offer concrete and minor suggestions. Those meta-commentary topics which directly request an edit or a review inspire

Sentiment change		
Feature	**Coefficients**	**Examples**
First person pronouns	.08**	I, me
Past tense	.04**	ago, did, talked
Positive emotion	.06**	great, nice, sweet
Reward	.06**	take, prize, benefit
Comparative modifiers	-.08*	significant amount, in particular
Quantitative modifiers	-.06*	many articles, most cases, vast majority
JJ NN (adjective+noun)	-.04*	relevant description, rare possibility
IN VBZ (preposition+verb)	-.04*	that is, in position
VBZ JJ (verb+adjective)	-.03*	seems correct, is only

Table 7: Analysis of the features significantly associated with a sentiment change, with *p<10^{-3}, **p<10^{-6}.

Editorial change		
Feature	**Coefficients**	**Examples**
Second person pronoun	.03*	you, your
Present tense	.01*	today, is, now
Work	.01*	job, work, source
Discrepancy	.01*	should, would
Superlative modifiers	.03*	great deal, great job, best place
IN (preposition)	-.04*	of, is, in, off, from
NNP (proper noun)	-.04*	axelboldt, gulag, wales
NN (noun)	-.04*	mistake, theory, problem

Table 8: Analysis of the features significantly associated with an editorial change, , with *p<10^{-3}, **p<10^{-6}.

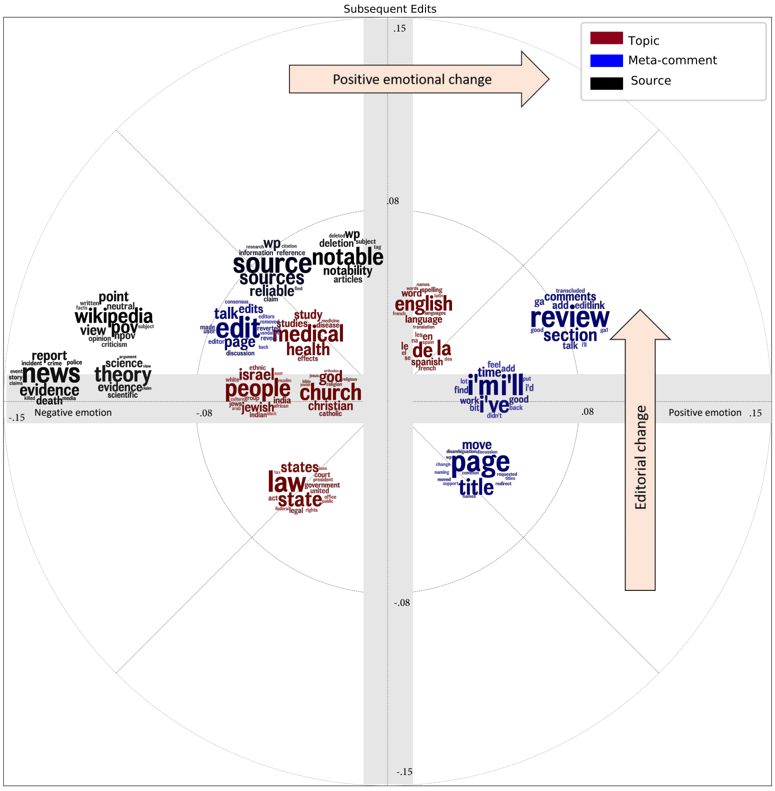

Figure 2: LDA topics correlated with emotional and editorial change. Topics are colored according to their theme; word size is proportional to word weight in the topic.

editorial change (.03< r < .06, p < .05). Finally, topics related to the source, i.e., about Wikipedia's guidelines, generate a more nuanced reaction. Topics related to evidentiality (*source, news, evidence*) and notability (*notable, articles, deletion*) are the strongest predictors of negative emotion (-.18<r<-.10, p < .05) but they generally lead to editorial changes (.03 < r < .08, p < .05).

9 Discussion and Limitations

An exploration of the WikiTalkEdit dataset suggests that strategies that elicit a positive emotional change may not affect editorial behavior. Negative responses should not be the only yardstick to measure the successful outcome of a conversation. Editorial changes occur when Editors use interpersonal language in talking about evidentiality and notability. However, these strategies are also associated with a negative emotional change. Despite the apparent negative feedback, referencing norms and sources is a successful strategy to prompt behavioral compliance. In related work, social influence through mentioning community norms was more effective than the Editor's status at achieving compliance on Wikipedia; however, the latter was an important predictor in a similar modeling task on Reddit (Althoff et al., 2014).

Although the findings would be correlational, there would be ways to establish cause and effect through a rigorous research design (Zhang et al., 2018). In some cases, the measurements may be thrown off if the replies to feedback are appreciative, but include some negative emotion words. Secondly, inordinately long or short feedback confounds the classifiers, but we expect that improvements in accuracy can be achieved by using differential attention models that focus on the emotions expressed in the first few words in the dialogic exchanges. Finally, we could encode the latent space with information about the type of editorial feedback (Yang et al., 2017), which would be helpful in predicting how the OP responds.

10 Conclusion and Future Applications

The WikiTalkEdit dataset offers insights that have important implications for understanding online disagreements and better supporting the Wikipedia community (Klein et al., 2019). We recommend the use of the WikiTalkEdit dataset to model the dynamics of consensus among multiple contributors. Scholars can also use the WikiTalkEdit dataset to address issues of quality, retention, and loyalty in online communities. For instance, the insights could shed light on how new OPs can be retained as sustaining Wikipedia contributors (Yang et al., 2017). Our exploratory analyses suggest that disagreements on Wikipedia arise over "errors": doubts that a given entry leaves no room for improvements. But errors serve a good faith purpose on Wikipedia by perpetuating participation and shared collective action (Nunes, 2011). The dataset would also be useful to understand how references are debated and interpreted as objective pieces of evidence (Luyt, 2015).

Acknowledgements: Supported by a Nanyang Presidential Postdoctoral fellowship and Templeton Religion Trust, grant TRT-0048. Thanks to Dr. Nicholas Palomares for their early feedback.

References

Tim Althoff, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. 2014. How to ask for a favor: A case study on the success of altruistic requests. In *Eigth International AAAI Conference on Web and Social Media (ICWSM 2015)*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

Moira Burke, Elisabeth Joyce, Tackjin Kim, Vivek Anand, and Robert Kraut. 2007. Introductions and requests: Rhetorical strategies that elicit response in online communities. In *Communities and Technologies 2007*, pages 21–39. Springer.

Marilyn J Chambliss and Ruth Garner. 1996. Do adults change their minds after reading persuasive text? *Written Communication*, 13(3):291–313.

Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. A computational approach to politeness with application to social factors. *arXiv preprint arXiv:1306.6078*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Amanda M Durik, M Anne Britt, Rebecca Reynolds, and Jennifer Storey. 2008. The effects of hedges in persuasive arguments: A nuanced analysis of language. *Journal of Language and Social Psychology*, 27(3):217–234.

Ivan Habernal and Iryna Gurevych. 2016. What makes a convincing argument? empirical analysis and detecting attributes of convincingness in web argumentation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1214–1223.

Craig R Hullett. 2005. The impact of mood on persuasion: A meta-analysis. *Communication Research*, 32(4):423–442.

Kokil Jaidka, Salvatore Giorgi, H Andrew Schwartz, Margaret L Kern, Lyle H Ungar, and Johannes C Eichstaedt. 2020. Estimating geographic subjective well-being from twitter: A comparison of dictionary and data-driven language methods. *Proceedings of the National Academy of Sciences*, 117(19):10165–10171.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431. Association for Computational Linguistics.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751. Association for Computational Linguistics.

Max Klein, Julia Kamin, and J Nathan Matias. 2019. 6 Ideas to Strengthen Wikipedia (s) with Citizen Behavioral Science. *citizensandtech.org/2019/11/research-summit-with-wikimedians*.

Adam DI Kramer, Jamie E Guillory, and Jeffrey T Hancock. 2014. Experimental evidence of massive-scale emotional contagion through social networks.

Proceedings of the National Academy of Sciences, 111(24):8788–8790.

Klaus Krippendorff. 1993. Conversation or intellectual lmperialism in comparing communication theories. *Communication Theory*, 3(3):252.

Siwei Lai, Liheng Xu, Kang Liu, and Jun Zhao. 2015. Recurrent convolutional neural networks for text classification. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, pages 2267–2273. AAAI Press.

Jürgen Lerner and Alessandro Lomi. 2019. The network structure of successful collaboration in wikipedia. In *Proceedings of the 52nd Hawaii International Conference on System Sciences*.

Ruoyun Lin and Sonja Utz. 2015. The emotional responses of browsing Facebook: Happiness, envy, and the role of tie strength. *Computers in human behavior*, 52:29–38.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Brendan Luyt. 2015. Debating reliable sources: writing the history of the vietnam war on wikipedia. *Journal of documentation*.

Vlad Niculae, Srijan Kumar, Jordan Boyd-Graber, and Cristian Danescu-Niculescu-Mizil. 2015. Linguistic harbingers of betrayal: A case study on an online strategy game. *arXiv preprint arXiv:1506.04744*.

Mark Nunes. 2011. *Error: glitch, noise, and jam in new media cultures*. Bloomsbury Publishing.

Ellie Pavlick and Joel Tetreault. 2016. An empirical analysis of formality in online communication. *Transactions of the Association for Computational Linguistics*, 4:61–74.

James W Pennebaker, Roger J Booth, and Martha E Francis. 2007. Linguistic inquiry and word count: LIWC.

Alec Radford. 2018. Improving language understanding by generative pre-training. *openai.com/blog/language-unsupervised*.

Bernard Rimé. 2007. The social sharing of emotion as an interface between individual and collective processes in the construction of emotional climates. *Journal of Social Issues*, 63(2):307–322.

Joachim Schulte. 1980. *Wittgenstein: an introduction*. SUNY Press.

Dhanya Sridhar, James Foulds, Bert Huang, Lise Getoor, and Marilyn Walker. 2015. Joint models of disagreement and stance in online debate. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 116–125.

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

Zhongyu Wei, Yang Liu, and Yi Li. 2016a. Is this post persuasive? Ranking argumentative comments in online forum. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, volume 2, pages 195–200.

Zhongyu Wei, Yandi Xia, Chen Li, Yang Liu, Zachary Stallbohm, Yi Li, and Yang Jin. 2016b. A preliminary study of disputation behavior in online debating forum. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 166–171.

Diyi Yang, Aaron Halfaker, Robert Kraut, and Eduard Hovy. 2016a. Edit categories and editor role identification in wikipedia. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1295–1299.

Diyi Yang, Aaron Halfaker, Robert Kraut, and Eduard Hovy. 2016b. Who did what: Editor role identification in wikipedia. In *Tenth International AAAI Conference on Web and Social Media (ICWSM 2016)*.

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.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *CoRR*, abs/1906.08237.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alexander J. Smola, and Eduard H. Hovy. 2016c. Hierarchical attention networks for document classification. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, page 1480–1489.

Justine Zhang, Jonathan P Chang, Cristian Danescu-Niculescu-Mizil, Lucas Dixon, Yiqing Hua, Nithum Thain, and Dario Taraborelli. 2018. Conversations gone awry: Detecting early signs of conversational failure. *arXiv preprint arXiv:1805.05345*.