Abstract

We investigate the novel task of online dispute detection and propose a sentiment analysis solution to the problem: we aim to identify the sequence of sentence-level sentiments expressed during a discussion and to use them as features in a classifier that predicts the DISPUTE/NON-DISPUTE label for the discussion as a whole. We evaluate dispute detection approaches on a newly created corpus of Wikipedia Talk page disputes and find that classifiers that rely on our sentiment tagging features outperform those that do not. The best model achieves a very promising F1 score of 0.78 and an accuracy of 0.80.

1 Introduction

As the web has grown in popularity and scope, so has the promise of collaborative information environments for the joint creation and exchange of knowledge (Jones and Rafaeli, 2000; Sack, 2005). Wikipedia, a wiki-based online encyclopedia, is arguably the best example: its distributed editing environment allows readers to collaborate as content editors and has facilitated the production of over four billion articles of surprisingly high quality (Giles, 2005) in English alone since its debut in 2001.

Existing studies of collaborative knowledge systems have shown, however, that the quality of the generated content (e.g. an encyclopedia article) is highly correlated with the effectiveness of the online collaboration (Kittur and Kraut, 2008; Kraut and Resnick, 2012): fruitful collaboration, in turn, inevitably requires dealing with the disputes and conflicts that arise (Kittur et al., 2007). Unfortunately, human monitoring of the often massive social media and collaboration sites to detect, much less mediate, disputes is not feasible.

In this work, we investigate the heretofore novel task of dispute detection in online discussions. Previous work in this general area has analyzed dispute-laden content to discover features correlated with conflicts and disputes (Kittur et al., 2007). Research focused primarily on cues derived from the edit history of the jointly created content (e.g. the number of revisions, their temporal density [Kittur et al., 2007; Yasseri et al., 2012]) and relied on small numbers of manually selected discussions known to involve disputes. In contrast, we investigate methods for the automatic detection, i.e. prediction, of discussions involving disputes. We are also interested in understanding whether, and which, linguistic features of the discussion are important for dispute detection.

Drawing inspiration from studies of human mediation of online conflicts (e.g. Billings and Watts (2010), Kittur et al. (2007), Kraut and Resnick (2012)), we hypothesize that effective methods for dispute detection should take into account the sentiment and opinions expressed by participants in the collaborative endeavor. As a result, we propose a sentiment analysis approach for online dispute detection that identifies the sequence of sentence-level sentiments (i.e. very negative, negative, neutral, positive, very positive) expressed during the discussion and uses them as features in a classifier that predicts the DISPUTE/NON-DISPUTE label for the discussion as a whole.

Consider, for example, the snippet in Figure 1 from the Wikipedia Talk page for the article on Philadelphia; it discusses the choice of a picture for the article’s “infobox”. The sequence of almost exclusively negative statements provides evidence of a dispute in this portion of the discussion.

Unfortunately, sentence-level sentiment tagging for this domain is challenging in its own right due to the less formal, often ungrammatical, language and the dynamic nature of online conversations. “Really, grow up” (segment 3) should presumably be tagged as a negative sentence as should the sarcastic sentences “Sounds good?” (in the same turn) and “congrats” and “thank you”
variants of Conditional Random Fields (Lafferty et al., 2001) and predicts sentiment at the turn-level, while our predictions are made for each sentence.

2 Data Construction: A Dispute Corpus

We construct the first dispute detection corpus to date; it consists of dispute and non-dispute discussions from Wikipedia Talk pages.

Step 1: Get Talk Pages of Disputed Articles.

Wikipedia articles are edited by different editors. If an article is observed to have disputes on its talk page, editors can assign dispute tags to the article to flag it for attention. In this research, we are interested in talk pages whose corresponding articles are labeled with the following tags: DISPUTED, TOTALLYDISPUTED, DISPUTED-SECTION, TOTALLYDISPUTED-SECTION, POV. The tags indicate that an article is disputed, or the neutrality of the article is disputed (POV).

We use the 2013-03-04 Wikipedia data dump, and extract talk pages for articles that are labeled with dispute tags by checking the revision history. This results in 19,071 talk pages.

Step 2: Get Discussions with Disputes.

Dispute tags can also be added to talk pages themselves. Therefore, in addition to the tags mentioned above, we also consider the “Request for Comment” (RFC) tag on talk pages. According to Wikipedia, RFC is used to request outside opinions concerning the disputes.

3609 discussions are collected with dispute tags found in the revision history. We further classify dispute discussions into three subcategories: CONTROVERSY, REQUEST FOR COMMENT (RFC), and RESOLVED based on the tags found in discussions (see Table 1). The numbers of discussions for the three types are 42, 3484, and 105, respectively. Note that dispute tags only appear in a small number of articles and talk pages. There may exist other discussions with disputes.

Additional Related Work.

Sentiment analysis has been utilized as a key enabling technique in a number of conversation-based applications. Previous work mainly studies the attitudes in spoken conversations among editors about the article content and allows editors to discuss the writing process, e.g. planning and organizing the content.
distinct speakers and 10 turns. 3609 discussions are randomly selected with this criterion. The average turn numbers for dispute and non-dispute discussions are 45.03 and 22.95, respectively.

3 Sentence-level Sentiment Prediction

This section describes our sentence-level sentiment tagger, from which we construct features for dispute detection (Section 4).

Consider a discussion comprised of sequential turns; each turn consists of a sequence of sentences. Our model takes as input the sentences \( x = \{x_1, \cdots, x_n\} \) from a single turn, and outputs the corresponding sequence of sentiment labels \( y = \{y_1, \cdots, y_n\} \), where \( y_i \in \{O, \mathcal{O} = \{NN, N, O, P, PP\} \). The labels in \( \mathcal{O} \) represent very negative (NN), negative (N), neutral (O), positive (P), and very positive (PP), respectively.

Given that traditional Conditional Random Fields (CRFs) (Lafferty et al., 2001) ignore the ordinal relations among sentiment labels, we choose isotonic CRFs (Mao and Lebanon, 2007) for sentence-level sentiment analysis as they can enforce monotonicity constraints on the parameters consistent with the ordinal structure and domain knowledge (e.g. word-level sentiment conveyed via a lexicon). Concretely, we take a lexicon \( \mathcal{M} = \mathcal{M}_p \cup \mathcal{M}_n \), where \( \mathcal{M}_p \) and \( \mathcal{M}_n \) are two sets of features (usually words) identified as strongly associated with positive and negative sentiment. Assume \( \mu_{(\sigma, w)} \) encodes the weight between label \( \sigma \) and feature \( w \), for each feature \( w \in \mathcal{M}_p \); then the isotonic CRF enforces \( \sigma \leq \sigma' \Rightarrow \mu_{(\sigma, w)} \leq \mu_{(\sigma', w)} \). For example, when we observe "totally agree" in the training data, the feature parameter for \( \mu_{(PP, totally agree)} \) is likely to increase. Similar constraints are defined on \( \mathcal{M}_n \).

Our lexicon is built by combining MPQA (Wilson et al., 2005), General Inquirer (Stone et al., 1966), and SentiWordNet (Esuli and Sebastiani, 2006) lexicons. Words with contradictory sentiments are removed. We use the features in Table 2 for sentiment prediction.

### Syntactic/Semantic Features

- unigram/bigram
- number of words all uppercased
- number of words

### Discourse Features

- initial uni-/bi-/tri-gram
- repeated punctuations
- hedging phrases collected from Parkas et al. (2010)
- number of negators

### Table 2: Features used in sentence-level sentiment prediction. Numerical features are first normalized by standardization, then binned into 5 categories.

#### Sentiment Features

- Baseline (Polarity) (1): a sentence is predicted as positive if it has more positive words than negative words, or negative if more negative words are observed. Otherwise, it is neutral. (2) Baseline (Distance) is extended from Hassan et al. (2010). Each sentence word is associated with the closest
In each column, **bold** entries (if any) are statistically significantly higher than all the rest. We also compare with an SVM and linear CRF trained with three classes (3-way). Our model based on the isotonic CRF produces significantly better results than all the other systems.

We choose number of turns, number of participants, average number of words in each turn as features. In addition, the frequency of revisions made during the discussion has been shown to be good indicator for controversial articles (Vuong et al., 2008), that are presumably prone to have disputes. Therefore, we encode the number of revisions that happened during the discussion as a feature.

### Sentiment Features

This set of features encode the sentiment distribution and transition in the discussion. We train our sentiment tagging model on the full AAWD dataset, and run it on the Wikipedia dispute corpus.

Given that consistent negative sentiment flow usually indicates an ongoing dispute, we first extract features from sentiment distribution in the form of number/probability of sentiment per type. We also estimate the sentiment transition probability $P(S_t \rightarrow S_{t+1})$ from our predictions, where $S_t$ and $S_{t+1}$ are sentiment labels for the current sentence and the next. We then have features as number/portion of sentiment transitions per type.

Features described above mostly depict the global sentiment flow in the discussions. We further construct a local version of them, since sentiment distribution may change as discussion proceeds. For example, less positive sentiment can be observed as dispute being escalated. We thus split each discussion into three equal length stages, and create sentiment distribution and transition features for each stage.

### Discussion Features

This type of feature aims to capture the structure of the discussion. Intuitively, the more turns or the more participants a discussion has, the more likely there is a dispute. Meanwhile, participants tend to produce longer utterances when they make arguments.

| Method                | Pos | Neg | Neutral |
|-----------------------|-----|-----|---------|
| Baseline (Polarity)   | 22.53 | 38.61 | 66.45 |
| Baseline (Distance)   | 33.75 | 55.79 | 88.97 |
| SVM (3-way)           | 44.62 | 52.36 | 80.84 |
| CRF (3-way)           | 56.28 | 56.37 | 89.41 |
| CRF (5-way)           | 58.39 | 56.30 | 90.10 |
| isotonic CRF          | 68.18 | 62.53 | 88.87 |

Table 3: F1 scores for positive and negative alignment on Wikipedia Talk pages (AAWD) using 5-fold cross-validation. In each column, **bold** entries (if any) are statistically significantly higher than all the rest. We also compare with an SVM and linear CRF trained with three classes (3-way). Our model based on the isotonic CRF produces significantly better results than all the other systems.

| Baseline (Polarity)   | Prec | Rec  | F1   | Acc  |
|-----------------------|------|------|------|------|
| Baseline (Random)     | 50.00 | 50.00 | 50.00 | 50.00 |
| Baseline (All dispute)| 50.00 | 100.00 | 66.67 | 50.00 |
| Logistic Regression   | 74.76 | 72.29 | 73.50 | 73.94 |
| SVM$_{Linear}$        | 69.81 | 71.90 | 70.84 | 70.41 |
| SVM$_{RBF}$           | 77.38 | 79.14 | 78.25 | 80.00 |

Table 4: Dispute detection results on Wikipedia Talk pages. The numbers are multiplied by 100. The items in **bold** are statistically significantly higher than others in the same column (paired-$t$ test, $p < 0.05$). SVM with the RBF kernel achieves the best performance in precision, F1, and accuracy.

### Results and Error Analysis

We experiment with logistic regression, SVM with linear and RBF kernels, which are effective methods in multiple text categorization tasks (Joachims, 1999; Zhang and J. Oles, 2001). We normalize the features by standardization and conduct a 5-fold cross-validation. Two baselines are listed: (1) labels are randomly assigned; (2) all discussions have disputes.
Sample sentences (sentiment in parentheses)
A: No, I sincerely plead with you. (N) If not, you are just wasting my time. (NN)
B: I believe Sweet’s proposal... is quite silly. (NN)
C: Tell you what. (NN) If you can get two other editors to agree... I will shut up and sit down. (NN)
D: But some idiot forging your signature claimed that doing so would violate. (NN)... Please go have some morning coffee. (O)
E: And I don’t like coffee. (NN) Good luck to you. (NN)
F: Was that all? (NN)... I think that you are in error... (N)

A: So far so confusing. (NN)... B: I can not see a rationale for the lanercase having its own article... (N) With Turkish Van being a miserable stub, there’s no such rationale for forking off a new article... (NN)... C: I’ve also copied your post immediately above to that article’s talk page since it is a great “nutshell” summary. (PP)
D: Err... how can the opposite be true... (N)
E: Thanks for this, though I have to say some of the facts floating around this discussion are wrong. (P)
F: Great. (PP)... Let’s make sure the article is clear on this. (O)

Figure 2: Sentiment flow for a discussion with unresolved dispute about the definition of “white people” (top) and a discussion with resolved dispute on merging articles about van cat (bottom). The labels {NN, N, O, P, PP} are mapped to {−2, −1, 0, 1, 2} in sequence. Sentiment values are convolved by using a Gaussian smoothing kernel, and then cubic-spline interpolation is conducted. Different speakers are represented by curves of different colors. Dashed vertical lines delimit turns. Representative sentences are labeled with letters and their sentiment labels are shown on the right. For unresolved dispute (top), we see that negative sentiment exists throughout the discussion. Whereas, for the resolved dispute (bottom), less negative sentiment is observed at the end of the discussion; participants also show appreciation after the problem is solved (e.g. E and F in the plot).

Table 5: Dispute detection results with different feature sets by SVM with RBF kernel. The numbers are multiplied by 100. Senti\_g represents global sentiment features, and Senti\_g\_{+1} includes both global and local features. The number in bold is statistically significantly higher than other numbers in the same column (paired-t test, p < 0.05), and the italic entry has the highest absolute value.

| Feature Set                  | Prec | Rec | F1  | Acc  |
|-----------------------------|------|-----|-----|------|
| Lexical (Lex)               | 78.36| 74.06| 76.51| 79.88|
| Topic (Top)                 | 68.44| 71.46| 69.92| 69.26|
| Discussion (Dis)            | 69.73| 76.14| 71.79| 71.54|
| Sentiment (Senti\_g\_{+1})  | 72.54| 69.52| 71.00| 71.60|
| Top + Dis                   | 68.49| 71.79| 70.10| 69.38|
| Top + Dis + Senti\_g\_{+1}  | 77.39| 78.36| 77.87| 77.74|
| Lex + Top + Dis + Senti\_g\_{+1} | 77.38| 79.14| 78.12| 80.00|

Main results for different classifiers are displayed in Table [4]. All learning based methods outperform the two baselines, and among them, SVM with the RBF kernel achieves the best F1 score and accuracy (0.78 and 0.80). Experimental results with various combinations of features sets are displayed in Table [5] As it can be seen, sentiment features obtains the best accuracy among the four types of features. A combination of topic, discussion, and sentiment features achieves the best performance on recall, F1, and accuracy. Specifically, the accuracy is significantly higher than all the other systems (paired-t test, p < 0.05).

After a closer look at the results, we find two main reasons for incorrect predictions. Firstly, errors from sentiment prediction get propagated into dispute detection. Due to the limitation of existing general-purpose lexicons, some opinionated dialog-specific terms are hard to catch. For example, “I told you over and over again...” strongly suggests a negative sentiment, but no single word shows negative connotation. Constructing a lexicon tuned for conversational text might further improve the performance. Secondly, some dispute discussions are harder to detect than the others due to different dialog structures. For instance, the recalls for dispute discussions of “controversy”, “RFC”, and “resolved” are 0.78, 0.79, and 0.86 respectively. We intend to design models that are able to capture dialog structures, such as pragmatic information, in the future work.

Sentiment Flow Visualization. We visualize the sentiment flow of two disputed discussions in Figure [2]. The plots reveal persistent negative sentiment in unresolved disputes (top). For the resolved dispute (bottom), participants show gratitude when the problem is settled.

5 Conclusion

We present a sentiment analysis-based approach to online dispute detection. We create a large-scale dispute corpus from Wikipedia Talk pages to study the problem. A sentiment prediction model based on isotonic CRFs is proposed to output sentiment labels at the sentence-level. Experiments on our dispute corpus also demonstrate that classifiers trained with sentiment tagging features outperform others that do not.

Acknowledgments We heartily thank the Cornell NLP Group, the reviewers, and Yiye Ruan for helpful comments. We also thank Emily Bender and Mari Ostendorf for providing the AAWD dataset. This work was supported in part by NSF grants IIS-0968450 and IIS-1314778, and DARPA DEFT Grant FA8750-13-2-0015. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of NSF, DARPA or the U.S. Government.
References

[Bender et al.2011] Emily M. Bender, Jonathan T. Morgan, Meghan Oxley, Mark Zachry, Brian Hutchinson, Alex Marin, Bin Zhang, and Mari Ostendorf. 2011. Annotating social acts: Authority claims and alignment moves in wikipedia talk pages. In Proceedings of the Workshop on Languages in Social Media, LSM ’11, pages 48–57, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Billings and Watts2010] Matt Billings and Leon Adam Watts. 2010. Understanding dispute resolution online: using text to reflect personal and substantive issues in conflict. In Elizabeth D. Mynatt, Don Schoner, Geraldine Fitzpatrick, Scott E. Hudson, W. Keith Edwards, and Tom Rodden, editors, CHI, pages 1447–1456. ACM.

[Esuli and Sebastiani2006] Andrea Esuli and Fabrizio Sebastiani. 2006. Sentiwordnet: A publicly available lexical resource for opinion mining. In Proceedings of the 5th Conference on Language Resources and Evaluation (LREC06), pages 417–422.

[Farkas et al.2010] Richárd Farkas, Veronika Vincze, György Móra, János Csirik, and György Szarvas. 2010. The conll-2010 shared task: Learning to detect hedges and their scope in natural language text. In Proceedings of the Fourteenth Conference on Computational Natural Language Learning — Shared Task, CoNLL ’10: Shared Task, pages 1–12, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Galley et al.2004] Michel Galley, Kathleen McKeown, Julia Hirschberg, and Elizabeth Shriberg. 2004. Identifying agreement and disagreement in conversational speech: use of Bayesian networks to model pragmatic dependencies. In ACL ’04: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, pages 669–709, Morristown, NJ, USA. Association for Computational Linguistics.

[Giles2005] G. Giles. 2005. Internet encyclopaedias go head to head. Nature, 438(7070):900–901.

[Hahn et al.2006] Sangyun Hahn, Richard Ladner, and Mari Ostendorf. 2006. Agreement/disagreement classification: Exploiting unlabeled data using contrast classifiers. In Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers, pages 53–56, New York City, USA, June. Association for Computational Linguistics.

[Hassan et al.2010] Ahmed Hassan, Vahed Qazvinian, and Dragomir Radev. 2010. What’s with the attitude?: Identifying sentences with attitude in online discussions. In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, EMNLP ’10, pages 1245–1255, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Hirschberg and Litman1993] Julia Hirschberg and Diane Litman. 1993. Empirical studies on the disambiguation of cue phrases. Comput. Linguist., 19(3):501–530, September.

[Joachims1999] Thorsten Joachims. 1999. Advances in kernel methods. chap. Making Large-scale Support Vector Machine Learning Practical, pages 169–184. MIT Press, Cambridge, MA, USA.

[Jones and Rafaeli2000] Q. Jones and S. Rafaeli. 2000. Time to split, virtually: discourse architecture and community building create vibrant virtual publics. Electronic Markets, 10:214–223.

[Kittur and Kraut2008] Aniket Kittur and Robert E. Kraut. 2008. Harnessing the wisdom of crowds in wikipedia: Quality through coordination. In Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work, CSCW ’08, pages 37–46, New York, NY, USA. ACM.

[Kittur et al.2007] Aniket Kittur, Bongwon Suh, Bryan A. Pendleton, and Ed H. Chi. 2007. He says, she says: Conflict and coordination in wikipedia. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI ’07, pages 453–462, New York, NY, USA. ACM.

[Kraut and Resnick2012] R. E. Kraut and P. Resnick. 2012. Building successful online communities: Evidence-based social design. MIT Press, Cambridge, MA.

[Lafferty et al.2001] John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Proceedings of the Eighteenth International Conference on Machine Learning, ICML ’01, pages 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

[Mao and Lebanon2007] Yi Mao and Guy Lebanon. 2007. Isotonic conditional random fields and local sentiment flow. In Advances in Neural Information Processing Systems.

[Rashmi Prasad and Webber2008] Alan Lee Eleni Miltysakaki Livio Robaldo Aravind Joshi Rashmi Prasad, Nikhil Dinesh and Bonnie Webber. 2008. The penn discourse treebank 2.0. In Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC’08), Marrakech, Morocco, may. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2008/.

[Sack2005] W. Sack. 2005. Digital formations: It and new architectures in the global realm. chap. Discourse architecture and very large-scale conversation, pages 242–282. Princeton University Press, Princeton, NJ USA.
[Stone et al.1966] Philip J. Stone, Dexter C. Dunphy, Marshall S. Smith, and Daniel M. Ogilvie. 1966. *The General Inquirer: A Computer Approach to Content Analysis*. MIT Press, Cambridge, MA.

[Vuong et al.2008] Ba-Quy Vuong, Ee-Peng Lim, Aixin Sun, Minh-Tam Le, Hady Wirawan Lauw, and Kuivy Chang. 2008. On ranking controversies in wikipedia: Models and evaluation. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, WSDM ’08, pages 171–182, New York, NY, USA. ACM.

[Wang et al.2011] Wen Wang, Sibel Yaman, Kristin Precoda, Colleen Richey, and Geoffrey Raymond. 2011. Detection of agreement and disagreement in broadcast conversations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2*, HLT ’11, pages 374–378, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Wilson et al.2005] Theresa Wilson, Janice Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT ’05, pages 347–354, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Yasseri et al.2012] Taha Yasseri, Róbert Sumi, András Rung, András Kornai, and János Kertész. 2012. Dynamics of conflicts in wikipedia. *CoRR*, abs/1202.3643.

[Yin et al.2012] Jie Yin, Paul Thomas, Nalin Narang, and Cecile Paris. 2012. Unifying local and global agreement and disagreement classification in online debates. In *Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis*, WASSA ’12, pages 61–69, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Zhang and J. Oles2001] Tong Zhang and Frank J. Oles. 2001. Text categorization based on regularized linear classification methods. *Inf. Retr.*, 4(1):5–31, April.