Identifying Sensible Participants in Online Discussions

Siddharth Jain
siddhajj@usc.edu
University of Southern California

Abstract

This paper investigates the problem of identifying participants in online discussions whose contribution can be considered sensible. Sensibleness of a participant can be indicative of the influence a participant may have on the course/outcome of the discussion, as well as other participants in terms of persuading them towards his/her stance. The proposed sensibleness model uses features based on participants' contribution and the discussion domain to achieve an F1-score of 0.89 & 0.78 for *Wikipedia: Articles for Deletion* and *4forums.com* discussions respectively.

1 Introduction

In contentious online discussions, people are very quick to classify other participants as being 'sensible' or not. What exactly this means is very hard to define. However, if one looks beyond the flippant 'anyone who agrees with me is sensible', it is possible to identify characteristics that tend to signal more thoughtful contributions. These include avoiding ad hominem attacks, making contributions that others respond favorably towards, obeying common rules of discourse, and so on. Sensibleness of a participant is quantified based on his/her contribution to the discussion, which is relevant to the discussion and reasoned in a way that is appealing to other participants.

In this paper, domain independent characteristics are identified and their stability is tested through human annotations to develop a classification system for determining sensibleness of participants in discussions on Wikipedia and 4forums.com. The proposed method leverages features obtained through argumentation mining. Domain specific characteristics are also incorporated in the analysis of the Wikipedia corpus.

2 Related Work

The pioneering work in argumentation mining is that of Moens (Moens et al., 2007), who addressed mining argumentation from legal documents. Recently, the focus has moved to mining user-generated content, such as online debates (Cabrio and Villata, 2012), discussions on regulations (Park and Cardie, 2014), and product reviews (Ghosh et al., 2014). Hasan (Hasan and Ng, 2014) use a probabilistic framework for argument recognition jointly with the related task of *stance* classification. Rosenthal (Rosenthal and McKeown, 2012) detect opinionated claims in online discussions in which author expresses a belief. They investigate the impact of features such as sentiment and committed belief on their system.

To date, almost no computational work has focused on the surface signals of "sense" in rhetoric. Danescu-Niculescu-Mizil (Danescu-Niculescu-Mizil et al., 2013) proposes a framework for identifying politeness. Although politeness seems an important aspect in identifying sensibleness, it is not mandatory. For example, the comment "I don't care how much you love the city. It cannot be on Wikipedia as it doesn't have enough coverage to satisfy Wikipedia policy." doesn't seem polite, though the author does seem sensible. Sun (Sun and Ng, 2012) propose a graph model to represent

Proceedings of The Fourth International Workshop on Natural Language Processing for Social Media, pages 41–47,
Austin, TX, November 1, 2016. ©2016 Association for Computational Linguistics

the relationship between online posts of one topic, in order to identify influential users. Tang (Tang and Yang, 2012) proposed a new approach to incorporate users' reply relationships to identify influential users of online healthcare communities. All these network based approaches determine the influence of a participant based on his/her centrality to the community/discussion and do not pay much attention to the specific content provided by the participants.

3 Corpus and Annotation

The corpus (Jain et al., 2014) for sensibleness annotation consists of 80 discussions from Wikipedia's Article for Deletion (AfD) discussion forum and 10 discussions from 4forums.com discussion forum. Sensibleness is highly dependent on the domain and nature of the discussion. Wikipedia discussions are goal-oriented: each participant tries to sway the decision of the discussion in their favor. Also, since Wikipedia pages should meet the requirements stated in their policies, one would expect the discussions to revolve around such policies. Therefore a criterion for people to be sensible in such discussions is that they appeal to authority in support of their arguments/claims. Additional criteria include not becoming emotional, avoiding tangents not relevant to the main topic, peer reviews, etc.

	Wikipedia	**4forums.com**
#Discussions	80	10
#participants	768	174
#Comments	1487	624
#Words	96138	51659

Table 1: Corpora stats.

In contrast, the discussions on 4forums.com are opinion-oriented, where participants primarily focus on presenting their own opinion and reasoning, but do not seriously consider that of others except to dispute it. In this domain, sensibleness analysis differs from the Wikipedia domain in several ways. First, expressing emotions may be considered sensible; second, tangential discussions that are not relevant to the main topic may be considered sensible if other participants follow.

Annotating sensibleness: Three annotators were asked to annotate the sensibleness of each participant in the discussions. The coding manual was created after several annotation rounds using different Wikipedia discussions through a process of refinement and consensus. Here are some of the questions the annotators seek to answer to determine sensibleness of a participant:

- "Does the participant sound reasonable and knowledgeable?"
- "How many positive/negative responses does the participant have?"
- "Does the participant start or get involved in tangential discussion?"
- "How much emotion does the participant express and what is the tone of it?"
- Does the participant mention Wikipedia policies? (For Wikipedia discussions only)

Each discussion is treated separately for annotation, i.e. a participant's sensibleness value for one discussion doesn't affect his/her sensibleness value for any other discussion. The possible values for sensibleness in the annotations are +1 (= sensible), -1 (= non-sensible), and 0 (= indeterminable). The annotation agreement score is kappa=0.73 using Fleiss' kappa (Fleiss, 1971) measure.

	Wikipedia	**4forums.com**
#Sensible	641	139
#Non-sensible	109	31
#Indeterminable	18	4

Table 2: Sensibleness distribution in the corpora.

Annotating claims: Analyzing the argumentation structure of participants' comments is an important aspect of the sensibleness model. For this analysis, Wikipedia discussions are annotated for claims and claim-links. A *claim* is defined as any assertion made in a discussion that the author intends the reader to believe to be true, and that can be disputed. A *claim-link* is defined as the causal/conditional dependency between claims. The same annotators performed this task, achieving an agreement score of kappa=0.76 for claim delimitation and kappa=0.81 for linkage.

4 Sensibleness Model

The classification model for sensibleness is created by extracting relevant features from participant's comments. Supervised machine learning is applied to determine the sensibleness value.

4.1 Argumentation Structure

The argumentation structure of the comments is an important aspect in determining sensibleness. For example, while "*This page violates Wikipedia policies*" and "*This page violates Wikipedia policies because it has no sources*" both express an opinion, the second is deemed more sensible because it provides a reason for the opinion. In contrast, "*Violent offenders can stay off our street*" presents an opinion that does not contain any claim and doesn't contribute anything significant toward the discussion. Therefore it can be considered non-sensible. The argumentation structure analysis is divided into three parts: *claim detection*, *claim delimitation*, and *claim-link detection*.

4.1.1 Claim Detection

Each sentence is classified as either having or not having a claim using several lexical features. The features include word n-grams(1-3), POS tag n-grams(1-3), and dependency triples (Marneffe et al., 2006). The classifier also uses generalized back-off features for n=grams and dependency triples as proposed by Joshi (Joshi and Penstein-Rosé, 2009). Similarly back-off features for lexical bigrams and trigrams are used. The motivation behind these features is the diversity of the topics that prevails in the discussions, which causes data sparsity with specific word combinations, which occur very infrequently. An SVM classifier with radial basis function is used to detect the sentences that express claims.

4.1.2 Claim Delimitation

Claim delimitation is useful since a sentence may contain multiple claims. The annotated sentences are pre-processed to add B_C, I_C, and O_C tags to each word, where B_C indicates a word starting a claim, I_C indicates a word inside a claim and O_C indicates a word outside any claim. Conditional Random Field (CRF) implemented in *CRFsuite*[1] is used to tag each word automatically using features like word n-grams(1-3), POS n-grams(1-3), and a binary feature for questions.

4.1.3 Claim-Link Detection

For claim-link detection, claim pairs are formed and determined whether they are linked. For each claim pair, features used include word and POS n-grams of the claims, word and POS unigrams for at most 5 words preceding and succeeding the claims, # of similar words between the claims, "claim distance" between the claims counting number of claims between them, and "sentence distance" between the claims counting how many sentences apart they are. An SVM classifier with radial basis function is used to detect claims that are linked.

From the argumentation structure analysis, the features extracted for the sensibleness analysis are: % of sentences made as claims, and % of claims linked to other claims.

4.2 Tangential Comments

Participants who tend to deflect from the main subject of the discussion are considered to be non-sensible. For each participant, each of his/her comments is categorized as tangential to the discussion or not. To quantify this, *itf-ipf*, a slightly modified version of *tf-idf*, is used to approximate tangentiality of any comment. For any tangential comment, the words used in the comment would be used relatively less than other words overall and would be used by relatively fewer participants. *tf* (term frequency) and *pf* (participant frequency: total number of participants who used the word in the discussion) are calculated and the *itf-ipf* value for each word *w* in a comment is computed as:

$$w_{itf-ipf} = \frac{1}{w_{tf}} * \log \frac{N}{w_{pf}} \qquad (1)$$

N = total number of participants in discussion.

Using the *itf-ipf* value for each word, the tangential quotient (*TQ*) for a comment (*C*) is calculated as:

$$TQ_C = \frac{\sum_{w \epsilon C} w_{itf-ipf}}{N_w} \qquad (2)$$

N_w = total number of words in comment.

The total *itf-ipf* value is divided by the total number of words to nullify the effect of the length of the comment. For Wikipedia discussions, if the value of

[1] http://www.chokkan.org/software/crfsuite/

TQ_C for a comment is more than 1.3 standard deviations from the average tangential quotient of the discussion ($\mu + 1.3\sigma$), the comment considered tangential. Similarly, for 4forums.com discussions, if the value of TQ_C for a comment is more than 1.5 standard deviations from the average tangential quotient of the discussion ($\mu + 1.5\sigma$), the comment considered tangential.

% of comments as tangential comments is used as one of the features for the sensibleness model.

4.3 Peer Reviews

Peer reviews provide an external opinion on the sensibleness of a participant. They therefore play a significant part in determining sensibleness of a participant, as a system with no domain knowledge of the discussion topic cannot verify the validity of their claims. For this analysis, all sentences that contain references to other participants are identified using *NLTK*[2] toolkit's NER (Named Entity Recognition) module. Second person pronouns in replies to other participants as reference are also identified. Next, the sentences that contain the reference are analyzed using NLTK's sentiment analysis module. If the sentence has non-neutral sentiment, then the polarity of the sentence is checked. If the polarity of the sentence is positive, then it is considered a positive review towards the participant who is referenced in the sentence. Similarly, if the polarity is negative, then it is considered a negative peer review.

# of positive reviews and # negative reviews are used as features for the sensibleness analysis.

4.4 Other Features

The following intuitive features are also part of the sensibleness analysis:

- % of sentences as questions: It can be a good strategy to ask questions related to the discussion, but asking too many questions can be considered as non-sensible.
- % of comments as personal attacks: This feature is useful for identifying participants who constantly attack others rather than presenting their own arguments. A similar method to that for peer reviews is used to identify comments that are targeted towards other participants and have negative polarity.

5 Experiments and Results

Weka[3] is used for all the classification tasks. The classifier for sensibleness model is trained using Wikipedia discussions over the features described in previous sections and is tested on both Wikipedia and 4forums.com discussions. For Wikipedia discussions, a domain specific feature of "Policy" is also incorporated based on the intuition that participants who mention Wikipedia policies in their comments are considered sensible. The best performing classifier for each of the argumentation structure experiment is used for the sensibleness model. The sensibleness model is compared with two baseline models ("Everyone" and "Bag of words") and several other models listed below:

- **Everyone**: Every participant is classified as sensible
- **Bag of words**: An SVM classifier with radial basis function trained on word n-grams(1-3)
- **Claims**: An SVM classifier with radial basis function trained on % of sentences containing claims
- **Claim-Links**: An SVM classifier with radial basis function trained on % of claims linked to other claims
- **Claims+Links**: An SVM classifier with radial basis function trained on % of sentences containing claims and % of claims linked to other claims
- **Tangential**: A participant is classified as sensible if he/she has less than 25% comments as tangential comments
- **Peer reviews**: A participant is classified as sensible if he/she has equal or more positive reviews than negative reviews
- **Questions**: An SVM classifier with radial basis function trained on % of sentences as questions
- **Personal attacks**: An SVM classifier with radial basis function trained on % of comments as personal attacks
- **Policy**: A participant is classified as sensible if he/she mentions Wikipedia policy in any of his/her comment. A small vocabulary is used to detect policy mentions in any comment.

McNemar's test is used to measure statistical significance. A significance difference in performance

[2] http://www.nltk.org/

[3] http://www.cs.waikato.ac.nz/ml/weka/

for $p < 0.01$ is depicted with ▲ (gain) and ▼ (loss) and for $p < 0.05$ is depicted with △ (gain) and ▽ (loss). 10-fold cross validation is used for testing Wikipedia models. After experimenting with several classifiers, the weighted precision, recall, and F1-score for the best classifier for each model is reported. SVM with radial basis function performs the best for both "Sensibleness" and "Sensibleness+Policy" models.

Model	**Precision**	**Recall**	**F1-score**
Everyone	0.70	0.83	0.76
Bag of words	0.71	0.80	0.75
Claims	0.78	0.83	0.80▲
Claim-Links	0.73	0.81	0.76
Claims+Links	0.81	0.85	0.82▲
Tangential	0.79	0.84	0.79▲
Peer reviews	0.76	0.82	0.78△
Questions	0.75	0.72	0.73
Personal attacks	0.73	0.76	0.75
Policy	0.77	0.80	0.78△
Sensibleness	0.86	0.88	0.87▲
Sensibleness+ Policy	0.88	0.90	0.89▲

Table 3: Sensibleness analysis for Wikipedia. Statistical significance is measured against "Everyone" model.

Since there are no discussion policies for 4forums.com, no corresponding models are created for it. The models trained on Wikipedia discussions are used to classify sensibleness on 4forums.com. Table 7 & Table 8 show the results for sensibleness analysis for Wikipedia and 4forums.com discussions respectively.

5.1 Error analysis

Looking at the errors made by the sensibleness model for Wikipedia discussions, we find that some are due to the inability of the argumentation structure detection system to identify claims for participants with very few sentences. Any participant with no identified claims is highly likely to be classified as non-sensible by the sensibleness model and therefore if the model is unable to detect claims then it is very likely that the model will classify such instances incorrectly. Using sensibleness models trained on Wikipedia discussions for sensibleness analysis of 4forums.com discussions fail mainly due to the difference in the argumentation structure of the two domains. Participants with lesser % claims/claim-links would be classified incorrectly on 4forums.com discussions.

Model	**Precision**	**Recall**	**F1-score**
Everyone	0.64	0.80	0.71
Bag of words	0.64	0.73	0.68
Claims	0.72	0.74	0.73▲
Claim-Links	0.65	0.74	0.69
Claims+Links	0.74	0.75	0.74▲
Tangential	0.74	0.71	0.72△
Peer reviews	0.69	0.78	0.72△
Questions	0.63	0.78	0.70
Personal attacks	0.69	0.71	0.70
Sensibleness	0.77	0.79	0.78▲

Table 4: Sensibleness analysis for 4forums.com. Statistical significance is measured against "Everyone" model.

6 Conclusions and Future Work

The work presented in this paper only scratches the surface of the problem of identifying sensible participants in discussions. Still, the success of the approach of counting some surface features to determine sensibleness is encouraging. The sensibleness analysis presented in this paper shows that argumentation structure and other intuitive features provide moderate accuracy for identifying sensible participants in online discussions. In future, we intend to follow up by using more subtle features identified by the annotators that are central to the model, such as identifying emotions and tones of comments. We hope this work provides an indication that it is possible to address this problem despite its difficulty and inspires other approaches.

Acknowledgments

The author thanks Prof. Eduard Hovy and Dr. Ron Artstein who provided insight and expertise that greatly assisted the research.

References

Rob Abbott, Marilyn Walker, Pranav Anand, Jean E. Fox Tree, Robeson Bowmani, and Joseph King. 2011. How can you say such things?!?: Recognizing disagreement in informal political argument. In *Proceedings of the Workshop on Languages in Social Media*, LSM '11, pages 2–11, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ehud Aharoni, Anatoly Polnarov, Tamar Lavee, Daniel Hershcovich, Ran Levy, Ruty Rinott, Dan Gutfreund, and Noam Slonim. 2014. A benchmark dataset for automatic detection of claims and evidence. In *in the Context of Controversial Topics", in Proceedings of the First Workshop on Argumentation and Computation, ACL 2014*.

Sinan Aral and Dylan Walker. 2012. Identifying influential and susceptible members of social networks. *Science*, 337(6092):337–341.

Shlomo Argamon, Moshe Koppel, James W. Pennebaker, and Jonathan Schler. 2009. Automatically profiling the author of an anonymous text. *Commun. ACM*, 52(2):119–123, feb.

Eytan Bakshy, Jake M. Hofman, Winter A. Mason, and Duncan J. Watts. 2011. Everyone's an influencer: Quantifying influence on twitter. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, WSDM '11, pages 65–74, New York, NY, USA. ACM.

Or Biran and Owen Rambow. 2011. Identifying justifications in written dialogs. In *Proceedings of the 2011 IEEE Fifth International Conference on Semantic Computing*, ICSC '11, pages 162–168, Washington, DC, USA. IEEE Computer Society.

Elena Cabrio and Serena Villata. 2012. Combining textual entailment and argumentation theory for supporting online debates interactions. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers - Volume 2*, ACL '12, pages 208–212, Stroudsburg, PA, USA. Association for Computational Linguistics.

Alexander Conrad, Janyce Wiebe, and Rebecca Hwa. 2012. Recognizing arguing subjectivity and argument tags. In *Proceedings of the Workshop on Extra-Propositional Aspects of Meaning in Computational Linguistics*, ExProM '12, pages 80–88, Stroudsburg, PA, USA. Association for Computational Linguistics.

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

Rob Ennals, Beth Trushkowsky, and John Mark Agosta. 2010. Highlighting disputed claims on the web. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 341–350, New York, NY, USA. ACM.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. 76(5):378–382.

Debanjan Ghosh, Smaranda Muresan, Nina Wacholder, Mark Aakhus, and Matthew Mitsui. 2014. Analyzing argumentative discourse units in online interactions. In *Proceedings of the First Workshop on Argumentation Mining*, pages 39–48, Baltimore, Maryland, June. Association for Computational Linguistics.

Kazi Saidul Hasan and Vincent Ng. 2014. Why are you taking this stance? identifying and classifying reasons in ideological debates. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 751–762, Doha, Qatar, October. Association for Computational Linguistics.

Siddharth Jain, Archna Bhatia, Angelique Rein, and Eduard Hovy. 2014. A corpus of participant roles in contentious discussions. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may. European Language Resources Association (ELRA).

Mahesh Joshi and Carolyn Penstein-Rosé. 2009. Generalizing dependency features for opinion mining. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, ACLShort '09, pages 313–316, Stroudsburg, PA, USA. Association for Computational Linguistics.

Namhee Kwon, Liang Zhou, Eduard Hovy, and Stuart W. Shulman. 2007. Identifying and classifying subjective claims. In *Proceedings of the 8th Annual International Conference on Digital Government Research: Bridging Disciplines & Domains*, dg.o '07, pages 76–81. Digital Government Society of North America.

Na Li and Denis Gillet. 2013. Identifying influential scholars in academic social media platforms. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '13, pages 608–614, New York, NY, USA. ACM.

M. Marneffe, B. Maccartney, and C. Manning. 2006. Generating typed dependency parses from phrase structure parses. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC-2006)*, Genoa, Italy, May. European Language Resources Association (ELRA).

Marie-Francine Moens, Erik Boiy, Raquel Mochales Palau, and Chris Reed. 2007. Automatic detection of arguments in legal texts. In *Proceedings of the 11th International Conference on Artificial Intelligence and Law*, ICAIL '07, pages 225–230, New York, NY, USA. ACM.

Raquel Mochales Palau and Marie-Francine Moens. 2009. Argumentation mining: The detection, classification and structure of arguments in text. In *Proceedings of the 12th International Conference on Artificial Intelligence and Law*, ICAIL '09, pages 98–107, New York, NY, USA. ACM.

Joonsuk Park and Claire Cardie. 2014. Identifying appropriate support for propositions in online user comments. In *Proceedings of the First Workshop on Argumentation Mining*, pages 29–38, Baltimore, Maryland, June. Association for Computational Linguistics.

Chaïm Perelman, 1979. *The New Rhetoric: A Theory of Practical Reasoning*, pages 1–42. Springer Netherlands, Dordrecht.

Sara Rosenthal and Kathleen McKeown. 2012. Detecting opinionated claims in online discussions. In *Sixth IEEE International Conference on Semantic Computing, ICSC 2012, Palermo, Italy, September 19-21, 2012*, pages 30–37.

Beiming Sun and Vincent TY Ng. 2012. Identifying influential users by their postings in social networks. In *Proceedings of the 3rd International Workshop on Modeling Social Media*, MSM '12, pages 1–8, New York, NY, USA. ACM.

Xuning Tang and Christopher C. Yang. 2012. Ranking user influence in healthcare social media. *ACM Trans. Intell. Syst. Technol.*, 3(4):73:1–73:21.

S. Toulmin. 1958. *The Uses of Argument*. Cambridge University Press, 2*nd* edition edition.