The Use of Text Similarity and Sentiment Analysis to Examine Rationales in the Large-Scale Online Deliberations

Wanting Mao
Department of Computer Science
The University of Western Ontario
London, ON, Canada
fiona.wt.mao@gmail.com

Lu Xiao
Faculty of Information & Media Studies
The University of Western Ontario
London, ON, Canada
lxiao24@uwo.ca

Robert Mercer
Department of Computer Science
The University of Western Ontario
London, ON, Canada
mercer@csd.uwo.ca

Abstract

To overcome the increasingly time consuming and potentially challenging identification of key points and the associated rationales in large-scale online deliberations, we propose a computational linguistics method that has the potential of facilitating this process of reading and evaluating the text. Our approach is novel in how we determine the sentiment of a rationale at the sentence level and in that it includes a text similarity measure and sentence-level sentiment analysis to achieve this goal.

1 Introduction

In an online deliberation situation where users join in and offer their opinions or suggestions, they are expected to provide the rationales that justify their standpoints in the deliberation. In the final decision making process, one expectedly needs to read through the content and weigh different key points and related rationales. Wikipedia Article for Deletion (AfD) deliberations represent one such example. In the Wikipedia community, any member can propose to delete an existing Wikipedia article. After an article is proposed to delete, a deliberation topic about the article is opened in the AfD forum. The community members can express their opinions (e.g., to keep or to delete the article) and provide their rationales within the specified time period. After that, a community member (often a Wikipedia administrator) closes the deliberation by making the final decision. Researchers have analyzed the Wikipedia AfD forum and have demonstrated that it presents a successful example of large-scale online deliberation by allowing many people to participate equally, encouraging people to deliberate, and producing rational and meaningful rationales (e.g., Schneider et al., 2012; Xiao & Askin, 2014). Wikipedia policy requires that the final decision about the article should be made based on the discussed rationales instead of the count of opinion votes. In practice many Wikipedia members who close the deliberations follow this policy, which implies the potential problem of representing the diverse rationales and identifying the influential ones in this context.

Generating the final decision of a large scale online deliberation can become a daunting task, as the amount of opinions and rationales in the deliberation content increases significantly. To facilitate this decision making process in large-scale online deliberations, we have developed a method that uses an existing text-to-text similarity measure and our developed sentence-level sentiment analysis algorithm to address this issue. Specifically, we first group participants' opinions according to the similarity measure, then we identify the positive, neutral, and negative sentiments suggested by the participants' rationales in each group, and finally we choose a representative rationale from each sentiment category in a group. With our method the diverse opinions and rationales are presented to the final decision maker through a representative set of the rationales, reducing the redundant information from the deliberation content so as to make the process of reading and evaluating the deliberation content more efficient.

Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pages 147–153,
Baltimore, Maryland, USA. June 27, 2014. ©2014 Association for Computational Linguistics

2 Related Work

2.1 Text Similarity

Recognizing the relation between texts (e.g., sentence to sentence, paragraph to paragraph) could help people better understand the context.

Text similarity can be interpreted as similarity between sentences, paragraphs, documents, etc. It has been used in various aspects in NLP such as information retrieval, text classification, and automatic evaluation. The most fundamental part is word similarity. We consider words to be similar in the following conditions: synonyms, antonyms, similar concept (e.g., red, green), similar context (e.g., doctor, hospital), and hyponym/hypernym relation (e.g., dog, pet).

WordNet, a word-to-word similarity library was developed by Pedersen et al. (2004), and has been widely used to compute the similarity at a coarser granularity (e.g., sentence-to-sentence similarity). Various methods to deal with text similarity have been proposed over the past decades. Mihalcea et al. (2006) proposed a greedy method to calculate the similarity score between two texts T1 and T2. Basically for each word in T1 (T2), the maximum similarity score to any word in T2 (T1) is used. The WordNet similarity can be used for assigning similarity scores between every pair of words in the two texts.

Rus and Lintean (2012) proposed an optimal method to compute text similarity based on word-to-word similarity. It is similar to the optimal assignment problem. Given a weighted complete bipartite graph (G = X È Y; X × Y), with weight w(xy) on edge xy, we need to find a matching from X to Y with a maximum total weight. Their results showed that the optimal method outperformed the greedy method in terms of accuracy and kappa statistics.

Other statistics-based algorithms are also developed to measure text similarity, e.g., the use of the Latent Dirichlet Allocation (LDA) model (Rus et al., 2013).

2.2 Sentiment Analysis

Sentiment analysis is meant to determine the polarity of a certain text, which can be positive, negative or neutral. Related academia and industries have been extensively investigating sentiment analysis methods over the last decade. While most of the early work in sentiment analysis is aimed at analyzing the polarity of customer reviews (e.g., Kim and Hovy, 2004; Hu and Liu, 2004; Turney, 2002), there is a proliferation in analyzing social media text (e.g., Balahur, 2013; Liebrecht et al., 2013; Bakliwal et al., 2012; Montejo-Raez et al., 2012) and online discussions (e.g., Sood et al., 2012a, 2012b).

Researchers have used a variety of approaches to detect the sentiment polarity of the given text. For example, in Kim and Hovy's system (2004) the sentiment region of the opinion is identified based on the extracted opinion holders and topics. The system combines the sentiments of the sentiment region and the polarity of the words to determine the polarity of the given text.

In Li and Wu's (2010) study, they interpreted the article as a sequence of key words and calculated the sentiment score of each key word based on the dictionary and its privative and modifier near it. In the analysis of the tweets, Balahur (2013) replaced the sentiment words and modifiers by sentiment labels (positive, negative, high positive and high negative) or modification labels (negator, intensifier or diminisher), and then applied Support Vector Machine Sequential Minimal Optimization (SVM SMO) to classify three different data sets.

Online discussions may have inappropriate use of language in some cases, which affects the online community management negatively. Sood et al. (2012a) proposed a multistep classifier by combining valence analysis and a SVM to detect insults and classify the insult object.

Researchers have also looked at the use of dependency tree-based method for sentiment classification. For instance, Nakagawa et al. (2010) used a probabilistic model of the information garnered from the dependency tree to determine the sentiment of a sentence. Rentoumi et al. (2010) combines word sense disambiguation, a rule-based system, and Hidden Markov Models (HMMs) to deal with figurative language (e.g. record-shattering day) in sentiment analysis. Moilanen and Pulman (2007) presented a compositional model for three-class (positive, negative, and neutral) phrase-level and sentence-level sentiment analysis. In their algorithm, each binary combination of a Head and Complement had a rule that determined which of the Head and Complement polarities dominated. In exceptional cases the rule inverts the polarity of the subordinate.

Socher et al. (2013) developed a Recursive Neural Tensor Network (RNTN) model. The authors showed that the accuracy obtained by RNTN outperformed a standard recursive neural network (RNN), matrix-vector RNN (MV-RNN), Naive Bayes (NB) and SVM. The advantage of

RNTN is especially evident when compared with the methods that only use bag of words (NB and SVM). This indicates the importance of using parse trees during sentiment analysis.

3 A Method for Identifying Representative Rationales in Online Deliberations

Our observation of the Wikipedia AfD forum suggests that one topic (e.g., notability) can appear multiple times in different rationales by different users. For example, two users' comments –"*Could be redirected to OpenXMA, the content of which isn't all that different from this article*" and "*Redirected to OpenXMA as suggested*"–are considered redundant.

The redundant information itself does not add a new perspective to final decision making. On the other hand, sometime the information about the same type of rationale represents different opinions about it. Here is one such example from an article's deletion discussion: "*redirecting the page to the lead actors future projects section will be cool*" and "*I don't think it is wise to redirect to the original film*".

To make the final decision making process more efficient, compared to human reading of all the deliberation content, we have developed a method that includes a text-to-text similarity measure and a sentence-level sentiment analysis algorithm. Specifically, we use text similarity to group the rationales according to the aspects they reflect so we can select some rationales from each aspect group instead of all of them. We note that although the rationales are redundant in showing the same aspect, the redundancy implies the importance of the aspect in the deliberation since they are used multiple times by users in justifying their opinions. So in our method, we record the number of members that proposed the same aspect assuming that this would indicate the level of importance of the aspect to some extent. .

With the rationales grouped according to the aspects that they involve (e.g., notability, credibility, etc), our method examines the sentiment polarity of each rationale in a group to further examine whether the rationale is positive or negative (e.g., the article is notable or not), or is neutral about the aspect. Then we can identify the representative rationales of an opinion by choosing those that have the highest similarity score in a group. In sum, the text-to-text similarity measure combined with our sentence-level sentiment analysis algorithm helps us identify the representative rationales of diverse opinions in an online deliberation. An overview of our method is shown in Figure 1.

We applied our method in analyzing Wikipedia Article for Deletion (AfD) deliberation content. Next we discuss how this method is used to analyze the content.

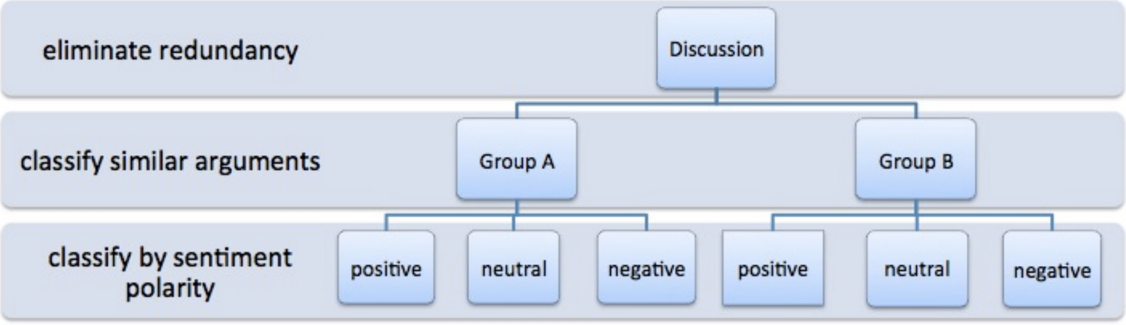

Figure 1. An overview of our method for identifying representative rationales from large-scale online deliberation

3.1 Text-to-Text Similarity Measure

In our study, we used SEMILAR, a semantic similarity toolkit (Rus et al., 2013), to measure We tested three similarity approaches provided in SEMILAR: optimum method based on WordNet, similarity based on Latent Semantic Analysis (LSA) and similarity based on Latent Dirichlet Analysis (LDA). We first extracted 80 pairs of sentences from the Wikipedia AfD forum and manually annotated them as similar or not. We then used these annotated results in measuring the accuracy of the three SEMILAR approaches. SEMILAR assigns a similarity score to each pair of sentences ranging from 0 to 1. To evaluate the accuracy of the three approaches, we identified a threshold to divide the result into two groups (i.e., similar and not similar). To do so, we computed the accuracy for 101 thresholds ranging from 0.00 to 1.00 with an interval of 0.01 to find the highest accuracy. Through this approach, we identified that the WordNet-based optimum method achieved the best accuracy of 76.3% at threshold 0.13. The other two methods achieved similar accuracy (76.3% and 75% respectively) but took more than double the time to process. Therefore, we chose the WordNet-based optimum method.

With this method, we have a similarity matrix that shows the similarity score between every pair of sentences in the discussion. We transform the similarity matrix to a dissimilarity matrix by transforming the similarity score *x* for two sentences to the distance between the sentences *1/x*. Then we used hierarchical clustering (Kaufman and Rousseeuw, 2009) to cluster the sentences into groups. To do so, we set the maximum allowed distance between two similar sentences to be 8 (i.e., the similarity score would be 0.125), and used the agglomerative approach to form the clusters. As a consequence, the sentences in the same group are related to a common theme.

3.2 Sentence-Level Sentiment Analysis

In our sentiment analysis algorithm, each word in a sentence is assigned a prior polarity based on an adapted MPQA Subjectivity Lexicon (Pedersen et al., 2004). Compared to the original Lexicon, this adapted one includes additional sentiment words that are important for the Wikipedia's AfD discussions (e.g., *notable).* Then, using the syntactic and dependency trees of the sentence, the algorithm calculates each word's current polarity score which can be affected by its children's polarity scores. Through this approach, the root's current polarity score becomes the sentence's polarity score.

The children's polarity scores can affect the parent's prior polarity score positively or negatively. The positive or neutral effect of the children's polarity scores is reflected through summing the children's polarity scores and then adding the sum to the parent's polarity score. The negative effect is reflected through summing the children's polarity scores and then multiplying the sum to the parent's polarity score. Because our algorithm only considers three sentiment situations: negative, positive, and neutral, it is the negation of the parent's prior polarity that affects the accuracy of our algorithm the most. Therefore, the core of our algorithm is a recursive method that examines different negation situations in the input sentence, starting from the leaf node of the sentence's dependency tree. We use this tree structure because it helps us detect the most of the negation situations:

1. I *agree* that the place is notable.
2. I *don't agree* that the place is notable. (Local Negation)
3. I *disagree* that the place is notable. (Predicate Negation)
4. *Neither* one of us agrees that the place is notable. (Subject Negation)
5. It is a *violation of* notability. (Preposition Negation)

However, there is one negation situation that cannot be detected from the syntactic structure of the sentence. For example, in the sentence "*the place is of indeterminable notability*", *notability* is a positive word, but as it is modified by a negative word *indeterminable* the phrase becomes negative. This negation case is called modifier negation. A negative modifier might also negate a negative word, such as *little damage*, *never fail*. However a negative modifier does not always negate the polarity of the phrase determined by the polarity of the related word. Instead, the phrase remains its prior polarity, e.g., *terribly allergic*.

It is also worth noticing that context affects the phrase polarity. Consider the phrase *original research* in our study context – the Wikipedia AfD forum. Because articles reporting original research violate Wikipedia's neutrality policy, the phrase *original research* in the deletion discussions should be considered to be negative.

As there is no straightforward way of determining whether or not a modifier negates the polarity of the word being modified, we decided to use machine learning methods to help classify the modifier negation cases. We considered the following modifier phrases in the study and at least one word in the phrase has to be a sentiment word:

- Noun modified by adjective
- Noun modified by noun
- Adjective modified by adverb
- Adverb modified by adverb
- Verb modified by adverb

We used six attributes to describe a two-word phrase: *first word token, second word token, first word polarity, second word polarity, first word part-of-speech (POS),* and *second word POS*. The machine learning algorithm is expected to predict the polarity of a word pair given these six attributes of the pair. To build our machine learning model, we obtained 961 two-word phrases from the AfD forum and annotated their polarities manually. They all follow the modifier negation combinations discussed earlier and at least one of the two words is a sentiment word. The selected phrases are balanced in terms of the number of positive, negative, and neutral cases represented in the data set. We then used Weka (Hall et al., 2009) to evaluate the performance of three machine learning algorithms with 10-fold cross validation: Naive Bayes, k-nearest neighbor (KNN) and decision tree. The results showed that the accuracy produced by KNN is the highest among the three methods. We further identified that when the k value is 1, the KNN performance is the best. Thus we selected the KNN method in detecting modifier negation in our method.

Figure 2 shows the calculated polarity score for the sentence "Neither one of us agrees that the place is notable".

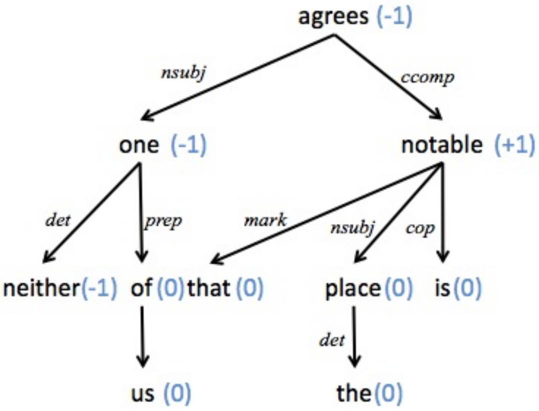

Figure 2. Polarity score on every node of the sentence' dependency structure

As shown in the figure, there are two positive words agree and notable and one negative word neither. If we simply use a bag of words approach and add the polarity scores together, we would get a result of positive. However, the negative word neither, being part of the subject, plays a dominant role in this sentence. Our algorithm is able to detect that negation influence: the root node is a verb and not neutral, so its current polarity score is the product of its prior polarity +1 multiplied by that of the node notable, which is also +1. Then because of the subject negation, the final polarity score of the root node is the multiplication of its current polarity score by the polarity score of the subject node, which is -1.

4 Evaluation and Discussion

To evaluate the performance of our sentiment polarity prediction algorithm, we randomly selected 236 sentences from the Wikipedia AfD forum and manually annotated their sentiment polarity. 83 sentences are annotated as positive, 102 as negative and 51 as neutral. With our algorithm that includes the machine learning process to detect modifier negations, the accuracy is 60.2%. In Socher et al.'s (2013) evaluation of their algorithm, 5-class (very negative, negative, neutral, positive, very positive) and 2-class (negative, positive) predictions of sentence-level sentiment analysis reached an accuracy of 45.7% and 85.4% respectively. We anticipate that the accuracy of their algorithm for 3-class prediction would be around 60%.

For sentence-level sentiment analysis, Moilanen and Pulman's algorithm obtained an accuracy of 65.6%. Our algorithm differs from Moilanen and Pulman in two ways: (1) the node-based computation is more general, i.e. for verbs, prepositions, and subjects it is a simple combination (multiplication or addition) of the subordinate nodes' polarities, and for local negation it is an inversion of the subordinate polarity; (2) a trained classifier serves two functions: it fulfills the role of determining the contextual information and it determines whether a modifier changes the polarity of what it modifies. .

5 Conclusion

Deliberation is a method of logical communication that rationalizes the process of reaching a decision. To reach the decision, people often need to weigh different opinions and rationales expressed in the deliberation. Given the proliferation of online platforms and communities for collective decision making and knowledge creation, online deliberation is becoming an increasingly important and common approach of engaging large numbers of people to participate in the decision making processes. One foreseen issue in such a context is the daunting tasks of reading through all the deliberation content, and identifying and evaluating diverse key points and related rationales.

Our study is interested in addressing the issue through a computational linguistic approach. We developed an approach that combines a text-to-text similarity technique with a sentence-level sentiment analysis method. The deliberation content is first divided into groups based on the similarity of texts, then within each group we use a recursive algorithm to examine the sentiment polarity of each sentence according to the identified similar topic to further classify the sentences into three groups: positive, neutral, and negative. Although not discussed in this paper, it is a simple step to identify the representative rationales of diverse opinions by choosing those that have the highest similarity score in each polarity group.

Acknowledgement

This project is partially supported by the Discovery program of The Natural Sciences and Engineering Research Council of Canada (NSERC).

References

Akshat Bakliwal, Piyush Arora, Senthil Madhappan-Nikhil Kapre, Mukesh Singh and Vasudeva Varma, Mining Sentiments from Tweets, Proceedings of the 3rd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis. 11–18, Jeju, Republic of Korea, 2012

Alexandra Balahur. Sentiment Analysis in Social Media Texts. WASSA 2013, page 120. 2013.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. The Weka data mining software: an Update. ACM SIGKDD Explorations Newsletter, 11(1):10–18, 2009.

Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge discovery and data mining, pages 168–177. ACM, 2004.

Leonard Kaufman and Peter J Rousseeuw. Finding groups in data: an introduction to cluster analysis, volume 344. John Wiley & Sons, 2009.

Soo-Min Kim and Eduard Hovy. Determining the sentiment of opinions. In Proceedings of the 20th international conference on Computational Linguistics, page 1367. Association for Computational Linguistics, 2004.

Nan Li, and Desheng Dash Wu. Using text mining and sentiment analysis for online forums hotspot detection and forecast. Decision Support Systems 48(2):354-368. 2010.

Christine Liebrecht, Florian Kunneman, and Antal van den Bosch (2013). The perfect solution for detecting sarcasm in tweets# not, Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, 29–37, Atlanta, Georgia, 2013.

Rada Mihalcea, Courtney Corley, and Carlo Strapparava. Corpus-based and knowledge-based measures of text semantic similarity. In AAAI, volume 6, pages 775–780, 2006.

Karo Moilanen and Stephen Pulman. Sentiment composition. In In Proceedings of Recent Advances in Natural Language Processing, pages 378 – 382, 2007.

Arturo Montejo-Raez, Eugenio Martiınez-Camara, M. Teresa Martin-Valdivia and L.Alfonso Urena-Lopez, Random Walk Weighting over SentiWordNet for Sentiment Polarity Detection on Twitter, Proceedings of the 3rd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis. 11–18, Jeju, Republic of Korea, 2012.

Tetsuji Nakagawa, Kentaro Inui, and Sadao Kurohashi. Dependency tree-based sentiment classification using CRFs with hidden variables, In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pages 786-794, 2010.

Ted Pedersen, Siddharth Patwardhan, and Jason Michelizzi. Wordnet:: Similarity: measuring the relatedness of concepts. In Demonstration Papers at HLT-NAACL 2004, pages 38–41. Association for Computational Linguistics, 2004.

Vassiliki Rentoumi, Stefanos Petrakis, Manfred Klenner, George A. Vouros, and Vangelis Karkaletsis. United we stand: Improving sentiment analysis by joining machine learning and rule based methods. In Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10), Valletta, Malta. pages 1089 – 1094, 2010.

Vasile Rus and Mihai Lintean. A comparison of greedy and optimal assessment of natural language student input using word-to-word similarity metrics. In Proceedings of the Seventh Workshop on Building Educational Applications Using NLP, pages 157–162. Association for Computational Linguistics, 2012.

Vasile Rus, Mihai Lintean, Rajendra Banjade, Nobal Niraula, and Dan Stefanescu. Semilar: The semantic similarity toolkit. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pages 163–168. 2013.

Jodi Schneider, Alexandre Passant, and Stefan Decker. Deletion discussions in Wikipedia: Decision factors and outcomes. In Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration, page 17. ACM, 2012.

Richard Socher, Alex Perelygin, Jean Y Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1631–1642, 2013.

Sara Owsley Sood, Elizabeth F. Churchill, and Judd Antin. Automatic identification of personal insults on social news sites. J. Am. Soc. Inf. Sci., 63: 270–285. 2012a.

Sara Sood, Judd Antin, and Elizabeth Churchill. 2012. Profanity use in online communities. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12). ACM, New York, NY, USA, 1481-1490, 2012b.

Peter D Turney. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pages 417–424. Association for Computational Linguistics, 2002.

Lu Xiao and Nicole Askin. What influences online deliberation? A Wikipedia study. Journal of the Association for Information Science and Technology, 65, pages 898–910, 2014