Is This Post Persuasive? Ranking Argumentative Comments in the Online Forum

Zhongyu Wei[12], **Yang Liu**[2] and **Yi Li**[2]
[1]School of Data Science, Fudan University, Shanghai, P.R.China
[2]Computer Science Department, The University of Texas at Dallas
Richardson, Texas 75080, USA
{zywei,yangl,yili}@hlt.utdallas.edu

Abstract

In this paper we study how to identify persuasive posts in the online forum discussions, using data from Change My View sub-Reddit. Our analysis confirms that the users' voting score for a comment is highly correlated with its metadata information such as published time and author reputation. In this work, we propose and evaluate other features to rank comments for their persuasive scores, including textual information in the comments and social interaction related features. Our experiments show that the surface textual features do not perform well compared to the argumentation based features, and the social interaction based features are effective especially when more users participate in the discussion.

1 Introduction

With the popularity of online forums such as idebate[1] and convinceme[2], researchers have been paying increasing attentions to analyzing persuasive content, including identification of arguing expressions in online debates (Trabelsi and Zaıane, 2014), recognition of stance in ideological online debates (Somasundaran and Wiebe, 2010; Hasan and Ng, 2014; Ranade et al., 2013b), and debate summarization (Ranade et al., 2013a). However, how to automatically determine if a text is persuasive is still an unsolved problem.

Text quality and popularity evaluation has been studied in different domains in the past few years (Louis and Nenkova, 2013; Tan et al., 2014; Park et al., 2016; Guerini et al., 2015). However, quality evaluation of argumentative text in the online forum has some unique characterisitcs. First, persuasive text contains argument that is not common in other genres. Second, beside the text itself, the interplay between a comment and what it responds to is crucial. Third, the community reaction to the comment also needs to be taken into consideration.

In this paper, we propose several sets of features to capture the above mentioned characteristics for persuasive comment identification in the online forum. We constructed a dataset from a sub-forum of Reddit[3], namely *change my view* (CMV)[4] . We first analyze the corpus and show the correlation between the human voting score for an argumentative comment and its entry order and author reputation. Then for the comment ranking task, we propose three sets of features including surface text features, social interaction based features and argumentation based features. Our experimental results show that the argumentation based features work the best in the early stage of the discussion and the effectiveness of social interaction features increases when the number of comments in the discussion grows.

2 Dataset and Task

2.1 Data

On CMV, people initiate a discussion thread with a post expressing their thoughts toward a specific topic and other users reply with arguments from the opposite side in order to change the initiator's mind. The writing quality on CVM is quite good since the discussions are monitored by moderators. Besides commenting, users can vote on different replies to indicate which one is more persuasive than others. The total amount of upvotes

[1]http://idebate.org/
[2]http://convinceme.net
[3]https://www.reddit.com
[4]https://www.reddit.com/r/changemyview

Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pages 195–200,
Berlin, Germany, August 7-12, 2016.

Thread #	1,785
Comment #	374,472
Comment # / Thread #	209.79
Author #	32,639
Unique Author # / Thread #	70.67
Delta Awarded Thread #	886 (49.6%)
Delta Awarded Comment #	2,056 (0.5%)

Table 1: Statistics of the CMV dataset.

minus the down votes is called *karma*, indicating the persuasiveness of the reply. Users can also give *delta* to a comment if it changes their original mind about the topic. The comment is then named *delta awarded comment* (DAC), and the thread containing a DAC is noted as *delta awarded thread*.

We use a corpus collected from CMV.[5] The original corpus contains all the threads published between Jan. 2014 and Jan. 2015. We kept the threads with more than 100 comments to form our experimental dataset[6]. The basic statistics of the dataset can be seen in Table 1.

Figure 1a shows the distribution of the karma scores in the dataset. We can see that the karma score is highly skewed, similar to what is reported in (Jaech et al., 2015). 42% of comments obtain a karma score of exactly one (i.e., no votes beyond the author), and around 15% of comments have a score less than one. Figure 1b and 1c show the correlation of the karma score with two metadata features, author reputation[7] and entry order, respectively. We can see the karma score of a comment is highly related to its entry order. In general, the earlier a comment is posted, the higher karma score it obtains. The average score is less than one when it is posted after 30 comments. Figure 1c shows that authors of comments with higher karma scores tend to have higher reputation on average.

2.2 Task

Tan et al. (2016) explored the task of mind change by focusing on delta awarded comments using their CMV data. However, the percentage of delta awarded comments is quite low, as shown in Table 1 (the percentage of comments obtained delta is as low as 0.5%). In addition, a persuasive comment is not necessarily delta awarded. It can be of high quality but does not change other people's mind. Our research thus uses the karma score of a comment, instead of delta, as the reference to represent the persuasiveness of the comment. Our analysis also shows that delta awarded comments generally have high karma scores (78.7% of DACs obtain a higher karma score than the median value in each delta awarded thread), indicating the karma score is correlated with the delta value.

Using karma scores as ground truth, Jaech et al. (2015) proposed a comment ranking task on several sub-forums of Reddit. In order to reduce the impact of timing, they rank each set of 10 connective comments. However, their setting is not suitable for our task. First, at the later stage of the discussion, comments posted connectively in terms of time can belong to different sub-trees of the discussion, and thus can be viewed or reacted with great difference. Second, as shown in Figure 1b, comments entered in later stage obtain little attention from audience. This makes their karma scores less reliable as the ground-truth of persuasiveness.

To further control the factor of timing, we define the task as ranking the first-N comments in each thread. The final karma scores of these N comments are used to determine their reference rank for evaluation. We study two setups for this ranking task. First we use information until the time point when the thread contains only these N comments. Second we allow the system to access more comments than N. Our goal is to investigate if we can predict whether a comment is persuasive and how the community reacts to a comment in the future.

3 Methods

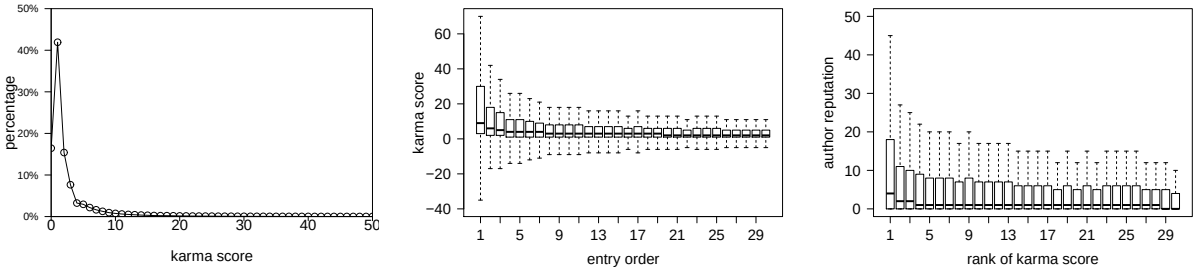
3.1 Ranking Model

A pair-wise learning-to-rank model (Ranking SVM (Joachims, 2002)) is used in our task. We first construct the training set including pairs of comments. In each pair, the first comment is more persuasive than the second one. Considering that two samples with similar karma scores might not be significantly different in terms of their persuasiveness, we propose to use a modified score to form training pairs in order to improve the learning efficacy. We group comments into 7 buckets based on their karma scores, [-∞, 0], (0, 1], (1, 5], (5, 10], (10, 20], (20, 50] and (50, +∞]. We then use the bucket number (0 - 6) of each comment

[5]The data was shared with us by researchers at the University of Washington.

[6]Please contact authors about sharing the data set.

[7]This is the number of deltas the author has received.



(a) Karma score distribution (b) Karma score vs entry order (c) Karma score vs author reputation

Figure 1: Karma value distributions in the CMV dataset.

Feature Category	Feature Name	Feature Description
Surface Text Features	length	# of the words, sentences and paragraphs in *c*.
	url	# of urls contained in *c*.
	unique # of words	# of unique words in *c*.
	punctuation	# of punctuation marks in *c*.
	unique # of POS	# of unique POS tags in *c*.
Social Interaction Features	tree_size	The tree size generated by *c* and *rc*.
	reply_num	The number of replies obtained by *c* and *rc*.
	tree_height	The height of the tree generated by by *c* and *rc*.
	Is_root_reply	Is *c* a root reply of the post?
	Is_leaf	Is *c* a leaf of the tree generated by *rc*?
	location	The position of *c* in the tree generated by *rc*.
Argumentation Related Features	connective words	Number of connective words in *c*.
	modal verbs	Number of modal verbs included in *c*.
	argumentative sentence	Number and percentage of argumentative sentences.
	argument relevance	Similarity with the original post and parent comment.
	argument originality	Maximum similarity with comments published earlier.

Table 2: Feature list (*c*: the comment; *rc*: the root comment of *c*.)

as its modified score. We use all the formed pairs to train our ranker. In order to be consistent, we use the first-N comments in the training threads to construct the training samples to predict the rank for the first-N comments in a test thread.

3.2 Features

We propose several key features that we hypothesize are predictive of persuasive comments. The full feature list is given in Table 2.

- **Surface Text Features**[8]: In order to capture the basic textual information, we use the comment length and content diversity represented as the number of words, POS tags, URLs, and punctuation marks. We also explored unigram features and named entity based features, but they did not improve system performance and are thus not included.

- **Social Interaction Features**: We hypothesize that if a comment attracts more social attention from the community, it is more likely to be persuasive, therefore we propose several social interaction features to capture the community reaction to a comment. Besides the reply tree generated by the comment, we also consider the reply tree generated by the root comment[9] for feature computing. The *tree size* is the number of comments in the reply tree. The position of *c* is its level in the reply tree (the level of root node is zero).

- **Argumentation Related Features**: We believe a comment's argumentation quality is a good indicator of its persuasiveness. In order to capture the argumentation related information, we propose two sub-groups of features based on the comment itself and the interplay between the comment and other comments in the discussion. **a) Local features:** we trained a binary classifier to classify sentences as argumentative and non-argumentative using features proposed in (Stab and Gurevych, 2014). We then use the number and percentage of argumentative sen-

[8]Stanford CoreNLP (Manning et al., 2014) was used to preprocess the text (i.e., comment splitting, sentence tokenization, POS tagging and NER recognition.).

[9]It is a comment that replies to the original post directly.

Approach	NDCG@1	NDCG@5	NDCG@10
random	0.258	0.440	0.564
author	0.382	0.567	0.664
entry-order	0.460	0.600	0.689
LTR$_{text}$	0.372	0.558	0.658
LTR$_{social}$	0.475†	0.650†	0.718†
LTR$_{arg}$	0.475†	0.652†	0.725†
LTR$_{text+social}$	0.494†	0.666†	0.733†
LTR$_{text+arg}$	0.485†	0.654†	0.729†
LTR$_{social+arg}$	0.502†‡	0.674†‡	0.740†
LTR$_{T+S+A}$	0.508†‡	0.676†‡	0.743†‡
LTR$_{all}$	**0.521**†‡	**0.685**†‡	**0.752**†‡

Table 3: Performance of first-10 comments ranking (*T+S+A*: the combination of the three sets of features we proposed; *all*: the combination of two meta-data features and our features; **bold**: the best performance in each column; †: the approach is significantly better than both metadata baselines ($p < 0.01$); ‡: the approach is significantly better than LTR approaches using a single category of features ($p < 0.01$).).

tences predicted by the classifier as features. Besides, we include some features used in the classifier directly (i.e. number of connective words[10] and modal verbs). **b) Interactive features:** for these features, we consider the similarity of a comment and its parent comment, the original post, and all the previously published comments. We use cosine similarity computed based on the term frequency vector representation. Intuitively a comment needs to be relevant to the discussed topic and possibly have some original convincing opinions or arguments to receive a high karma score.

4 Experimental Results

We use 5-fold cross-validation in our experiments. Normalized discounted cumulative gain (NDCG) score (Järvelin and Kekäläinen, 2000) is used as the evaluation metric for our First-N comments ranking task. In this study, N is10.

4.1 Experiment I: Using N Comments for Ranking

Table 3 shows the results for first-10 comments ranking using information from only these 10 comments. As shown in Figure 1, metadata features, entry order and author's reputation are correlated with the karma score of a comment. We thus use these two values as baselines. We also include the performance of the random baseline for comparison[11]. For our ranking based models (*LTR*$_*$), we compare using the three sets of features described in Section. 3.2 (noted as *text*, *social* and *arg* respectively), individually or in combination. We report NDCG scores for position 1, 5 and 10 respectively. The followings are some findings.

- Both metadata based baselines generate significantly[12] better results compared to the random baseline. Baseline *entry-order* performs much better than *author*, suggesting that the entry order is more indicative for the karma score of a comment.

- The surface text features are least effective among the three sets of features, and the performance using them is even worse than the two metadata baselines. This might be because the general writing quality of the comments in CMV is high because of the policy of the forum. Therefore, the surface text features we used are not very discriminative for comment ranking. A further analysis of features in this category shows that *length* is the most effective feature.

- Argumentation based features have the best performance among the three categories. Its performance is significantly better than surface text features, consistent with our expectation that argumentation related features are useful for persuasiveness evaluation. Our additional experiments show that *interactive features* are more effective than *local features*. This might be because the argumentation features and models we use are not perfect. Future research is still needed to better represent argumentation information in the text.

- When combining two categories of features, the performance of the ranker increases consistently. The performance can be further improved by combining all the three categories of features we proposed (the improvement compared to using a single feature category is significant). The best results are achieved by *LTR*$_{all}$, i.e., combining two metadata features and features we proposed.

[10] We constructed a list of connective words including 55 entries (e.g., because, therefore etc.).

[11] The performance of random baseline is high because of the tie of reference karma scores.

[12] Significance is computed by two tailed t-test.

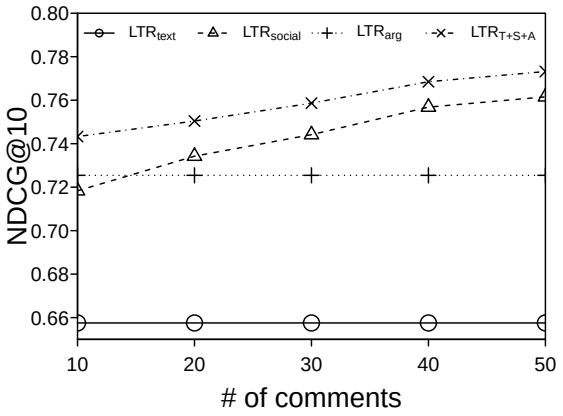


Figure 2: Results using various number of comments in the thread for ranking.

4.2 Experiment II: Using Varying Numbers of Comments for Ranking

With the evolving discussion, there will be more comments joining the thread providing more information for social interaction based features. In order to show the impact of different features at different discussion stage, we conduct another experiment by ranking first-10 comments with varying numbers of comments in the test thread for feature computing. The result of the experiment is shown in Figure 2. The performance of LTR_{text} and LTR_{arg} remain the same since their feature values are not affected by the new coming comments. The performance of LTR_{social} increases consistently when the number of comments grows, and it outperforms LTR_{arg} when the number of comments is more than 20. LTR_{T+S+A} has always the best performance, benefiting from the combination of different types of features.

5 Related Work

Our work is most related to two lines of work, including text quality evaluation and research on Reddit.com.

Text quality: Text quality and popularity evaluation has been studied in different domains in the past few years. Louis and Nenkova (2013) implemented features to capture aspects of great writing in science journalism domain. Tan et al. (2014) looked into the effect of wording while predicting the popularity of social media content. Park et al. (2016) developed an interactive system to assist human moderators to select high quality news. Guerini et al. (2015) modeled a notion of euphony and explored the impact of sounds on different forms of persuasiveness. Their research focused on the phonetic aspect instead of language usage.

Reddit based research: Reddit has been used recently for research on social news analysis and recommendation (e.g., (Buntain and Golbeck, 2014)). Researchers also analyzed the language use on Reddit. Jaech et al. (2015) studied how language use affects community reaction to comments in Reddit. Tan et al. (2016) analyzed the interaction dynamics and persuasion strategies in CMV.

6 Conclusion

In this paper, we studied the impact of different sets of features on the identification of persuasive comments in the online forum. Our experiment results show that argumentation based features work the best in the early stage of the discussion, while the effectiveness of social interaction based features increases when the number of comments in the thread grows.

There are three major future directions for this research. First, the approach for argument modeling in this paper is lexical based, which limits the effectiveness of argumentation related features for our task. It is thus crucial to study more effective ways for argument modeling. Second, we will explore persuasion behavior of the argumentative comments and study the correlation between the strength of the argument and different persuasion behaviors. Third, we plan to automatically construct an argumentation corpus including pairs of arguments from two opposite sides of the topic from CMV, and use this for automatic disputing argument generation.

Acknowledgments

We thank the anonymous reviewers for their detailed and insightful comments on this paper. The work is partially supported by DARPA Contract No. FA8750-13-2-0041 and AFOSR award No. FA9550-15-1-0346. Any opinions, findings, and conclusions or recommendations expressed are those of the authors and do not necessarily reflect the views of the funding agencies. We thank Trang Tran, Hao Fang and Mari Ostendorf at University of Washington for sharing the Reddit data they collected.

References

Cody Buntain and Jennifer Golbeck. 2014. Identifying social roles in reddit using network structure. In *Proceedings of the Companion Publication of the 23rd International Conference on World Wide Web Companion*, pages 615–620.

Marco Guerini, Gözde Özbal, and Carlo Strapparava. 2015. Echoes of persuasion: The effect of euphony in persuasive communication. *arXiv preprint arXiv:1508.05817*.

Kazi Saidul Hasan and Vincent Ng. 2014. Why are you taking this stance? identifying and classifying reasons in ideological debates. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 751–762.

Aaron Jaech, Vicky Zayats, Hao Fang, Mari Ostendorf, and Hannaneh Hajishirzi. 2015. Talking to the crowd: What do people react to in online discussions? In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 2026–2031.

Kalervo Järvelin and Jaana Kekäläinen. 2000. IR evaluation methods for retrieving highly relevant documents. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 41–48.

Thorsten Joachims. 2002. Optimizing search engines using clickthrough data. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 133–142.

Annie Louis and Ani Nenkova. 2013. What makes writing great? first experiments on article quality prediction in the science journalism domain. *Transactions of the Association for Computational Linguistics*, 1:341–352.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of the Association for Computational Linguistics System Demonstrations*, pages 55–60.

Deokgun Park, Simranjit Sachar, Nicholas Diakopoulos, and Niklas Elmqvist. 2016. Supporting comment moderators in identifying high quality online news comments. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*.

Sarvesh Ranade, Jayant Gupta, Vasudeva Varma, and Radhika Mamidi. 2013a. Online debate summarization using topic directed sentiment analysis. In *Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining*, page 7.

Sarvesh Ranade, Rajeev Sangal, and Radhika Mamidi. 2013b. Stance classification in online debates by recognizing users' intentions. In *Proceedings of Special Interest Group on Discourse and Dialogue*, pages 61–69.

Swapna Somasundaran and Janyce Wiebe. 2010. Recognizing stances in ideological on-line debates. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 116–124.

Christian Stab and Iryna Gurevych. 2014. Identifying argumentative discourse structures in persuasive essays. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 46–56.

Chenhao Tan, Lillian Lee, and Bo Pang. 2014. The effect of wording on message propagation: Topic- and author-controlled natural experiments on twitter. *arXiv preprint arXiv:1405.1438*.

Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2016. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. *arXiv preprint arXiv:1602.01103*.

Amine Trabelsi and Osmar R Zaıane. 2014. Finding arguing expressions of divergent viewpoints in online debates. In *Proceedings of the 5th Workshop on Language Analysis for Social Media (LASM)@ EACL*, pages 35–43.