Automatic Classification of Persuasive Arguments

Joshua Berlinski
jxb2648@rit.edu

Gleb Promokhov
gxp5819@rit.edu

Abstract

Public web forums allow for massive online debate, especially on the community platform Reddit. Persuasive arguments can be found on the sub-community /r/ChangeMyView[1], which encourages sharing views and discourse in a moderated public forum. To determine if there are any similarities between persuasive comments that were successful in changing a user's view we organized and labeled sets of argument examples, and found valuable features for classifying novel arguments through language modeling. Our model results saw 6% improved accuracy over the baseline, concluding that there are identifiable stylistic features in persuasive arguments.

1 Introduction

Being able to persuade someone to change their opinion has been a valued skill throughout the history of discourse. With words and discussion people have been able to convince others to adopt new views about the world. The ability to persuade is often influenced by the persuader's social standing, context of the discussion, and many more variables that make it difficult to isolate how the the language used affects that ability. Thousands of lines of discourse, free from these influences, would have to be collected to conduct an effective analysis of rhetorical language.

The on-line forum /r/ChangeMyView (CMV) provides an ideal corpus to conduct such an analysis. The anonymous, online format removes much of the social context during discourse. As of conducting this study CMV has had 470,082 user participate in an average of 60 discussions started per day. CMV has rules requiring minimum word counts for discussion, maintaining an encouraging discussion space and requiring user participation. In addition to this the community is heavily moderated so most spam and irrelevant content is quickly removed. This renders CMV a community with a high volume of quality data to research persuasive language free from the problems earlier described.

CMV works by having original posters (OPs) post a view and their rationale, inviting commenters to write dissenting views and to convince them to change their view. The OP and commenters often reply to each other and have a dialog before the OP decides that certain comments should be awarded a 'delta point' for having convinced them to change their view. These 'delta points' are logged onto a related sub-reddit /r/DeltaLog for record keeping.

This paper will be focusing on the stylistic and lexical differences between arguments to understand persuasive language separate from the content being argued. This paper will also examine the influence of discussing particular topics on the persuasiveness of an argument.

2 Prior Work

Previous work on this subject (Tan, et. al., 2016), notes that classifying persuasive arguments is a task that humans have difficulty with and may not be possible to generalize to all discourse. Their work also used CMV as their corpus, but we used improved classification techniques and larger dataset. Unlike the Tan, et. al., paper we found that including topic features greatly affects classification
accuracy.

3 Dataset

Certain attributes of CMV we found very conducive to our study. High user participation in many different debates on different topics ensure that our style analysis is not constrained to a specific subject matter or few user's writing styles. The online format does introduce the problem of having text written by bots, but minus a few used to run the forum CMV is explicitly written by humans only.

Users posting an view in CMV are required to show effort in their rationale, enforced through

[1] https://reddit.com/r/ChangeMyView

minimum characters limits (500+) and manual moderation removing 'low effort' content. Users are encouraged to be open to changing their view, to actively participate in discussion and avoid accusations, sarcasm, jokes and general negativity. These rules steer CMV's users towards high quality, serious discourse[2].

Comments are awarded a 'delta point' if they were successful in changing any part of a user's view. These labels made the task of finding negative and positive examples for the classification very straightforward. We found it fair to assume that these comments were more persuasive than those without delta points. We also decided that not only should the awarded comment's text be considered, but all comments by that comment's author up to the awarded comment should be considered persuasive as well.

3.1 Corpus

Our corpus consist of submissions on CMV made between 04/29/2017 and 11/02/2017, collected on 11/03/2017.

Table 1: Corpus sizes

	# comments
Training	4113
Testing	1029

Each post was recorded with its author, time of submission, and other metadata. Comments made on posts were organized into a tree, preserving the 'back-and-forth' discussion made between users on a post. Example in Fig. 1

3.2 Comment Pairs

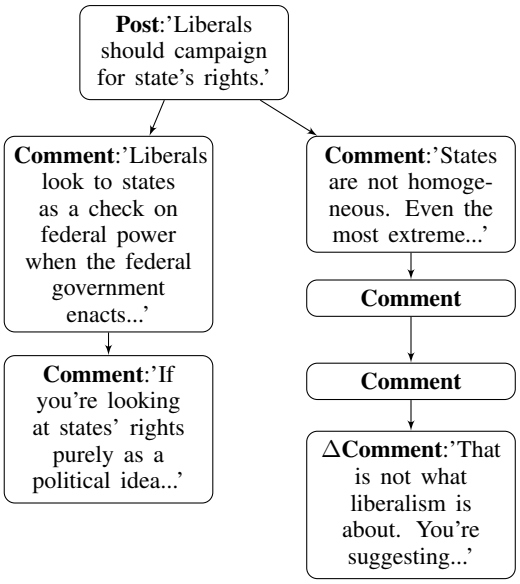
We extracted pairs of comments from each post where one comment received a delta point from OP and one that did not as positive and negative examples for classification. These comment pairs also have at least 13% Jaccard similarity in content words.

$$\text{Jaccard}(A, B) = \frac{|A \cap B|}{|A \cup B|} \quad (1)$$

With content word sets from both comments as *A* and *B*, this controls the notion that a comment received a delta point for its content, not its style. 13% was slightly above average of all comment similarities. We collected 6,000 pairs to train our classifier.

[2]A full list of rules can be found on the CMV wiki

Figure 1: Example Comment Tree



4 Model Features

For the classification task, the extracted features can be put into two categories. First, there are features extracted from the body of each comment. Second we assigned each comment a topic from set of topics generated from word clusters.

4.1 Body Features

The first features extracted from the comment bodies is the distribution of part-of-speech (POS) tags. These were collected by using the POS tagger included in the Python package `nltk` (Loper and Bird, 2002). POS tags provide a indication to the overall word choice and style of each comment. These tag features were recorded as the ratio between the count of the tag to total words in the comment. The next features included is the counts of positive and negative words using the lexicon by Hu and Liu (2004). To avoid problems with sentiment, any words between 'not' and any punctuation were prefixed with 'not_', and their sentiment was flipped (Polanyi and Zaenen, 2004; Kennedy and Inkpen, 2005). The sentiment of an argument should have an obvious effect on it's persuasive ability. As comments on Reddit may contain Markdown formatting, the count of italicized, bolded, and block-quoted text were included as features. These formatting features were included to capture some of the visual emphasis employed

in arguments. Block-quoted text captures when a user employs another user's argument as an example, either to highlight or debate it specifically. We also found that many arguments were supported by the use of hyperlinks. To account for this, features for the proportion of links that have the `.com` domain, the `.edu` domain, and that contain `.pdf`. These features are only present if links are present in the comment, otherwise they are zero. Links tend to include citations and references supporting a user's argument and possibly added to it's persuasiveness.

4.2 Topic Features

In addition to stylistic features we wanted to examine how the subject being argued affected if the argument was awarded a delta point or now. Topics clusters were generated using Dirichlet allocation (Blei et. al., 2003) topic modeling.The set of topics and their word clusters can be found in Table 4 in Appendix A.

5 Classification Task

Table 2: 5 Most influential features

Feature	Odds Ratio
POS Count: Interjection	5.92×10^{-5}
POS Count: Modal	915.75
POS Count: Wh-Determiner	822.8
POS Count: To	261.21
POS Count: Comp. Adj.	178.28

To classify CMV arguments, these features were used to train a logistic regression classifier. Since some features had high correlation with others, regression techniques that take feature structure into account were considered, such as SRIG (Yu and Liu, 2016), and the group LASSO (Meier, et. al., 2008). However, for ease of interpretation, features that had high correlation with others were removed, and the LASSO (Tibshirani, 1996) is used. The LASSO was chosen due to shrinkage property and the desirable selection property. Other regression methods with these properties such as the Elastic Net (Zou and Hastie, 2005) were experimented with, but these methods were outperformed by the LASSO.

The LASSO is a regression technique that seeks regression coefficients by minimizing the residual sum of the squared error subject to the L_1 norm of the coefficients being less than some LASSO parameter $t > 0$. That is, given a regression problem of the form

$$y_i = \sum_{j=1}^{p} \beta_j x_{ij} + \epsilon_i, \tag{2}$$

where y_i is the output variable for the i-th observation, β_j is the j-th regression coefficient, x_{ij} is the value of the j-th feature for the i-th observation, and ϵ is normally distributed error, the LASSO minimizes

$$\sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2, \quad \sum_{j=1}^{p} |\beta_j| < t. \tag{3}$$

For this task, the LASSO implementation in the Python package `scikit-learn` (Pedregosa, et. al., 2011) is used.

In order to evaluate the model, the original dataset was split into a training set with 80% of the comments and a validation set with the remaining 20%. The training set was further split into another 80% training and 20% testing for model tuning. In order to retain consistency in the topics given from the topic model, cross validation was not employed. After the model was tuned to yield the best results on the test set, it was then retrained on the full training set for the final test on the validation set.

Table 3: Confusion Matrix for Classification Task

		Predicted		
		$\neg\Delta$	Δ	**total**
Actual	$\neg\Delta$	280	234	514
	Δ	209	306	515
	total	489	540	

6 Results and Discussion

Accuracies reported were: 50% majority class baseline, 55.2% without topic features and 56.9% with topics.

As shown by the odds ratios in Table. 7 the high value of POS tags indicates that persuasive arguments have a specific word choice and style. High value of topic features indicate that users of CMV hold variably malleable views on different subjects.

An observation is that comparative adjectives are noticeably more present in persuasive arguments than superlative ones. Statements such as 'X is better, worse, larger, ... than Y' are more effective than stating 'X is the best'. Weighing the difference between two subjects is much more persuasive than absolute statements.

Modal verbs such as 'can, may, might, could' are more present in persuasive arguments. Modal verbs are used to indicate possibility; showing that persuasive arguments tend to offer (not absolutize) opinions and ask questions about the subject.

We also found that possessive pronouns, in phrases such as 'In my opinion', are much less effective at persuading. Introducing subjective personal views and vague anecdotes do not lead to effective arguments.

High value of including links, which tend to be citations and references, to support an argument make it more persuasive than otherwise.

In terms of topic features we found within topics dealing with sexuality (Topic 3 in Table. 4) users of CMV are very stubborn in their views. Users were more malleable with their views in topics 8 and 9, which deal with government spending and race issues.

7 Conclusion

Given our corpus we were able to correlate writing styles to their persuasiveness. Use of conjunctions and comparisons are more persuasive than personal or polarizing statements. Our analyses of persuasive language is of course only a small subset of all the possible conclusions we can draw from the CMV corpus. Much more exploration can be done into understanding the semantics behind the arguments, such as detecting effective syllogisms and logical conclusions. More work can be done towards consider the 'back-and-forth' in the discourse and examine how it plays a part in argument persuasiveness. We also considered including use of rhetorical language, words such as 'because', 'therefore', 'for example...', as features as well.

It is important to recognize that these features are shown to be significant with CMV comments, where users tend to write in a more conversational tone in a relaxed, anonymous format. Using our model to classify a corpus of user comments, forum posts, blogs and newspaper articles used in another study on argumentation (Habernal and Gurevych., 2017) we found minimal improvement over the baseline. More exploration will have to be done to analyze more formal argument structures.

Our corpus and code is available at `https://github.com/glebpro/nlptermproject2017`.

References

Blei, D., Ng, A., and Jordan, M. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research.* 3(Jan):993–1022.

Habernal, I., and Gurevych I. 2017. Argumentation Mining in User-Generated Web Discourse. *Computational Linguistics*, 43(1):125–179

Hu, M. and Liu, B. 2004 Mining and summarizing customer reviews. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*.

Kennedy A. and Inkpen D. 2005 Sentiment classification of movie and product reviews using contextual valence shifters. *Proceedings of the Workshop on the Analysis of Informal and Formal Information Exchange during Negotiations*, Ottowa, Ontario, Canada.

Loper, E. and Bird, S. 2002 NLTK: The Natural Language Toolkit. *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1*, 63–70.

Meier, L., van de Geer, S. and Bühlmann, P. 2008. The group lasso for logistic regression. *Journal of the Royal Statistical Society B*70:53–71.

Pedregosa, F., et. al. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830

Polanyi, L. and Zaenen, A. 2004 Contextual valence shifters *Exploring Attitute and Affect in Text: Theories and Applications (AAAI Spring Symposium Series).*

Tan, C., Niculae, V., Danescu-Niculescu-Mizil, C. and Lee, L. 2016. Winning Arguments: Interaction Dynamics and Persuasion Strategies in Good-faith Online Discussions. *Proceedings of WWW.*

Tibshirani, R. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society B*, 58(1):267–288.

Yu, G. and Liu, Y. 2016. Sparse Regression Incorporating Graphical Structure Among Predictors. *Journal of the American Statistical Association*, 111(514):707–720.

Zou, H. and Hastie, T. 2005. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society B*, 67(2):301–320.

Appendix A

Table 4: Topic word clusters

Topic	Name	Top 10 words
0	Gen. Life	*life, com, wiki, reddit, child, marriage, human, harm, like, wikipedia*
1	Politics	*trump, country, people, countries, vote, party, political, american, think, media*
2	Games	*game, food, time, like, games, new, ai, data, best, problem*
3	Sexuality	*gender, sex, sexual, person, gay, trans, body, male, women, woman*
4	Work	*money, need, people, cost, make, 10, time, work, better, company*
5		*people, like, don, just, women, know, think, men, say, ve*
6	Gen. Argument	*people, think, believe, don, argument, point, say, right, wrong, evidence*
7		*people, just, don, think, like, really, good, want, make, things*
8	Gov. Spending	*government, people, free, public, taxes, right, pay, military, state, business*
9	Race	*people, white, culture, black, racist, society, think, social, person, race*

Table 5: Full odds ratios

Feature	Odds Ratio
Interjection	5.9246946799471379e-05
Verb, Past Participle	0.019755523045652732
Wh-adverb	0.04285056789492029
Verb, sing, present	0.061764010273682016
Proper noun	0.14121389510143348
Verb, past	0.15210469962227346
Personal Pronoun	0.15439434933723242
Wh-pronoun	0.37569972474440455
Verb	0.41764117158361019
Topic 0	0.50510995216177457
Topic 3	0.5173609889609726
Topic 6	0.78402952380337354
Topic 2	0.85514875079765351
Topic 5	0.86045523988906281
Topic 4	0.8711109432426295
Topic 7	0.90488809987427821
Topic 1	0.92928416677830372
Block quotes	0.94537776943900442
Italic text	1.0008291732813517
Topic 8	1.0138484712637177
Positive Sentiment	1.0143272038057729
Bold text	1.0403242762666387
.com links	1.271146593385075
Verb, 3rd person pres.	1.6418100484061395
.pdf links	3.0437037755810961
.edu links	3.2894772187959651
Possessive ending	13.324969947247608
Verb, present participle	15.06923942984321
Comparative adjective	178.27973549864248
To	261.21440284600942
Wh-determiner	822.80066253560528
Modal	915.75471588262315