Tell me who you are, I'll tell whether you agree or disagree: Prediction of agreement/disagreement in news blogs

Fabio Celli and **Evgeny A. Stepanov** and **Giuseppe Riccardi**
Signals and Interactive Systems Lab
Department of Information Engineering and Computer Science
University of Trento, via Sommarive 5, Trento, Italy
{fabio.celli,evgeny.stepanov,giuseppe.riccardi}@unitn.it

Abstract

In this paper we address the problem of the automatic classification of agreement and disagreement in news blog conversations. We analyze bloggers, messages and relations between messages. We show that relational features (such as replying to a message or to an article) and information about bloggers (such as personality, stances, mood and discourse structure priors) boost the performance in the classification of agreement/disagreement more than features extracted from messages, such as sentiment, style and general discourse relation senses. We also show that bloggers exhibit reply patterns significantly correlated to the expression of agreement or disagreement. Moreover, we show that there are also discourse structures correlated to agreement (expansion relations), and to disagreement (contingency relations).

1 Introduction

Threaded discussions in on-line social media are asynchronous multiparty conversations that concur to the formation of opinions and shared knowledge which influence decision makers. Bloggers who participate in these conversations usually express their opinions, defend their stances and gain or lose consensus with their text messages. These conversations contain many layers of information such as sentiment [Strapparava and Mihalcea, 2008], humor [Reyes *et al.*, 2012], and Agreement/Disagreement

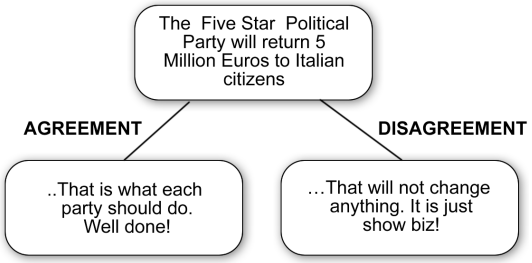

Figure 1: Example of agreement and disagreement.

Relations (henceforth ADRs) [Wang and Cardie, 2014] (see Figure 1 for an example). In this paper we address the problem of extracting ADRs from news blog conversations. There are two possible tasks: ADR detection (finding the messages that contain personal positions) and ADR classification (classifying messages as agreeing or disagreeing with previous messages). Here we address ADR classification and experiment with message, blogger and relational-level features and their combinations.

The paper is structured as follows: in Section 2 we provide an overview of previous work in the field and a definition of ADRs, from Section 3 to 6, we describe the data set, the annotation, the experimental settings and discuss the results.

2 Related Work and Definitions

Previous work on ADRs in asynchronous conversations can be divided into three areas: definition of ADRs, collection and annotation of corpora and prediction of ADRs' polarity.

In [Bender *et al.*, 2011] ADRs are considered

as relationships among bloggers expressed at message level with a post or turn text unit. They collected the AAWD corpus of Wikipedia talk pages and manually annotated with ADRs and authority claims. The reported inter-annotator reliability is k=0.5. In [Walker *et al.*, 2012] ADRs are defined as Quote-Response message pairs and triplets. These pairs and triplets are linked by the structure of the thread, where each message is a reply to its parent and is about the same topic. They collected the IAC corpus [Walker *et al.*, 2012] of political debates in English (about 2700 authors, 11k threads) extracted from *4forums.com* and annotated with ADRs by means of Amazon Mechanical Turk, obtaining inter-annotator reliability of α=0.62. In [Andreas *et al.*, 2012] ADRs are defined between pairs of sentences within messages in a parent/child relation. In their definition, ADRs have a type ("agree", "disagree" or "none") and a mode ("direct" or "indirect", "response" or "paraphrase"). They annotated sentence pairs in a corpus of LiveJournal and Wikipedia with 3 classes ("agree", "disagree", "not applicable"). The reliability between two annotators is k=0.73. In [Celli *et al.*, 2014] ADRs are defined as a function that maps pairs of bloggers' messages to polarity values between 1 ("agree") and -1 ("disagree"). They collected a corpus of news blogs conversations in Italian (CorEA corpus). The reported inter-annotator reliability is k=0.58 on 3 classes ("agree", "disagree", "not applicable") and k=0.87 on 2 classes ("agree", "disagree").

In [Wang and Cardie, 2014], the authors addressed ADRs classification between text segments corresponding to one or several sentences on the IAC and AAWD corpora. The authors observed that it is easier to classify agreement than disagreement in the AAWD corpus, while the contrary is true in the IAC corpus.

3 Dataset

The CorEA corpus [Celli *et al.*, 2014] is used for the experiments throughout the paper. As mentioned in the previous section, the corpus is the collection of news blogs in Italian and consists of asynchronous conversations from 27 news articles on different topics ranging from politics to gossip. The corpus contains 2,887 messages (135K tokens). The average number of messages per conversation is 106.4.

The corpus has been labeled by two annotators with three labels: "agreement", "disagreement" and "not applicable" (henceforth "NA"). Messages are annotated with a "NA" label, if they satisfy one or both of the following conditions: a) **message is not clear**, if the annotator cannot find or commit to the relation between parent and child messages (e.g. the child message contains only URLs or is not referred to its parent); b) **message contains mixed agreement**, if in the child message there are conflicting or ambiguous cues triggering agreement and disagreement. This includes cases such as conflicting opinions in the child message about one or more statements in the parent message. If the message does not fall under the cases specified above, the ADR in the child message is evaluated with respect to the parent as "agree" (1) or "disagree" (-1). The distribution of labels in the corpus is 31% agreement, 34% disagreement, and 35% NA.

4 Features

For the experiments on ADR classification, we exploit the features already present in the data and enriched them with new features at the level of messages, bloggers and parent-child relations (relational features).

Message-level Features (107). *Discourse Features* (8) are frequency counts and ratios (% from total) of the four top-level relation senses from Penn Discourse Treebank (PDTB) [Prasad *et al.*, 2008]: Comparison, Contingency, Expansion, and Temporal. They are extracted for explicit discourse relations (signaled by connectives such as *but*, *however*, *when*, etc.) using lexical context classifier of [Riccardi *et al.*, 2016]; and a connective sense classifier trained on Italian LUNA Corpus [Dinarelli *et al.*, 2009].*Sentiment Polarity Features* (2) are text-length normalized sums of the polarized words extracted using OpeNER lexicon (http://www.opener-project.eu/), and their discretisation into positive, neutral, and negative classes. *Stylometric Features* (97) are basic text statistics (4) such as word count, vocabulary size, average word length; frequency-based features (2) such as frequency of hapax legomena; measures of lexical richness (16) based on word count, vocabulary size, and word-frequency spectrum such as mean word frequency, type-token ratio, entropy, Guiraud's R, Honore's H, etc. [Tweedie and Baayen, 1998]; and word length ratios (30) for 1-30 character long words. The feature set also includes character-based ratios (45) for character classes (e.g. punctuation,

white space, etc.) and individual characters (e.g. '!', 'a', etc.). Additionally, we include the number of message likes and replies (2).

Relational-level Features (4). These features are generated using child and parent messages (or the article as parent). They include word2vec [Mikolov *et al.*, 2013] cosine similarity between parent and child messages, boolean feature to indicate whether a parent is an article or another message, and two boolean features for matches and mismatches between topics and sentiment polarities expressed in two messages.

Blogger-level Features (22). Blogger-level features are the personality types (5), self-assessed blogger mood priors (5); the aggregation (sums and averages) of the message-level discourse (8) features; blogger's stance (1), and blogger's topic per message ratio (1). Personality types are defined by the Five Factor Model: extroversion, emotional stability/neuroticism, agreeableness, conscientiousness, openness to experience. These features have been automatically predicted exploiting linguistic cues from the collection of all messages of single bloggers. The accuracy of the prediction, evaluated on an Italian Facebook dataset [Celli, 2013], is 65%. Mood priors encoded in CorEA are: indignation, disappointment, worry, amusement and satisfaction. Stance is the sum of the polarity of messages of a blogger.

5 Experiments and Results

As it was already stated, in this paper we address the problem of classification of ADRs as pairs of parent-child messages being in agreement or disagreement relation. Thus, the problem is case as a binary classification task; as opposed to the 3-way classification including "NA" relations or a two-step hierarchical ADR detection-classification task. For the experiments, we have balanced the data and partitioned it into training and testing as 66% and 33%. Since some blogger level features are aggregations of message-level features, the data was split by alphabetically sorting the messages by bloggers' names. Support Vector Machine classifier with linear kernel from Weka is used as learning algorithm.

The results on ADR classification using message, blogger-level and relational-level features and their combinations are reported in Table 1. Similar to the observation of [Wang and Cardie, 2014] for English on AAWD, we observe that classification performance for agreement is higher than disagreement for Italian as well. With respect to feature groups, we observe that blogger and relational features outperform the bag of words baseline. The best performance is obtained using relational feature only, followed by the blogger-level features. In order to evaluate the contributions of individual features, we have performed correlation analysis. Table 2 reports the ranking of the features highly correlated with agreement and disagreement labels. We also observe that bloggers who reply to the article tend to agree with its content, and this can be seen as a result of the quality of the information of the article and the credibility of news. In debate corpora (e.g. IAC) such a tendency is not observed. Moreover, bloggers that get more replies are the ones that disagree the most with others, and this can be explained with the fact that disagreement generates a debate. It is also interesting to note that extroversion is correlated to disagreement and conscientiousness to agreement. With respect to discourse structure, we observe that contingency and comparison rela-

settings	agree (F1)	disagree (F1)	both (acc)
majority baseline	0.500	0.500	0.500
bag of word baseline	0.550	0.624	0.590
message	0.555	0.554	0.550
blogger	0.634	0.568	0.601
relational	**0.726**	**0.684**	**0.705**
message+blogger	0.618	0.560	0.589
message+relational	0.711	0.675	0.693
blogger+relational	**0.726**	**0.684**	**0.705**
all	0.659	0.629	0.644

Table 1: Result of the classification of ADRs using different combinations of message features, blogger features and relational features. We use 66% training, 33% test split a Support Vector Machine as classifier (Weka SMOreg), F1 and accuracy (acc) as evaluation metrics.

Corr	feat. type	feature	Class
0.418	relational	article as parent	A
0.265	blogger	reply ratio	D
0.205	blogger	topic-message ratio	A
0.183	blogger	expansion	A
0.147	message	ratio of 2-char words	D
0.146	blogger	conscientiousness	A
0.127	message	! marks ratio	D
0.126	blogger	contingency	D
0.121	blogger	extroversion	D
0.106	blogger	comparison	D
0.105	message	replies count	D

Table 2: Ranking of the features highly correlated with agreement (A) and disagreement (D) (Pearson's correlation with $p - value < 0.001$).

tions tend to be used for expressing disagreement, while expansion relations are mainly used to express agreement. Moreover, this is complemented by the fact that bloggers in agreement with others tend to address more topics in a single message. Among the other observations, we notice that exclamation marks and short words are strong cues for disagreement.

6 Conclusion

In this paper we addressed the problem of classification of message-pairs from online conversations into agreement and disagreement relations. We have demonstrated that blogger-level and relational-level features outperform the message-level features, such as sentiment polarity and style. Through correlation analysis we have studied how agreement and disagreement relations are expressed. We have observed that there are discourse structures underlying the expression of agreement and disagreement relations in social media. The methodology presented in this paper is useful for the automatic analysis of online social media conversations. The future work includes the detection of agreement/disagreement relations and their exploitation for conversation summarisation.

Acknowledgements

The research leading to these results has received funding from the European Union – Seventh Framework Programme (FP7/2007–2013) under grant agreement 610916: SENSEI.

References

[Andreas *et al.*, 2012] Jacob Andreas, Sara Rosenthal, and Kathleen McKeown. Annotating agreement and disagreement in threaded discussion. In *LREC*, 2012.

[Bender *et al.*, 2011] Emily M Bender, Jonathan T Morgan, Meghan Oxley, Mark Zachry, Brian Hutchinson, Alex Marin, Bin Zhang, and Mari Ostendorf. Annotating social acts: Authority claims and alignment moves in wikipedia talk pages. In *Proceedings of WLSM*, pages 48–57. ACL, 2011.

[Celli *et al.*, 2014] Fabio Celli, Giuseppe Riccardi, and Arindam Ghosh. Corea: Italian news corpus with emotions and agreement. In *Proceedings of CLIC-it 2014*, pages 98–102, 2014.

[Celli, 2013] F Celli. *Adaptive Personality recognition from Text*. Lambert Academic Publishing, 2013.

[Dinarelli *et al.*, 2009] Marco Dinarelli, Silvia Quarteroni, Sara Tonelli, Alessandro Moschitti, and Giuseppe Riccardi. Annotating spoken dialogs: from speech segments to dialog acts and frame semantics. In *Proceedings of EACL Workshop on the Semantic Representation of Spoken Language*, Athens, Greece, 2009.

[Mikolov *et al.*, 2013] Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *HLT-NAACL*, 2013.

[Prasad *et al.*, 2008] R. Prasad, N. Dinesh, A. Lee, E. Miltsakaki, L. Robaldo, A. Joshi, and B. Webber. The penn discourse treebank 2.0. In *Proc. of LREC*, 2008.

[Reyes *et al.*, 2012] Antonio Reyes, Paolo Rosso, and Davide Buscaldi. From humor recognition to irony detection: The figurative language of social media. *Data & Knowledge Engineering*, 74:1–12, 2012.

[Riccardi *et al.*, 2016] Giuseppe Riccardi, Evgeny A. Stepanov, and Shammur Absar Chowdhury. Discourse connective detection in spoken conversations. In *Proc. of ICASSP*, 2016.

[Strapparava and Mihalcea, 2008] Carlo Strapparava and Rada Mihalcea. Learning to identify emotions in text. In *Proceedings of the 2008 ACM Symposium on Applied Computing*, SAC '08, pages 1556–1560, New York, NY, USA, 2008. ACM.

[Tweedie and Baayen, 1998] Fiona J. Tweedie and R. Harald Baayen. How variable may a constant be? measures of lexical richness in perspective. *Computers and the Humanities*, 32(5):323–352, 1998.

[Walker *et al.*, 2012] Marilyn A Walker, Jean E Fox Tree, Pranav Anand, Rob Abbott, and Joseph King. A corpus for research on deliberation and debate. In *LREC*, 2012.

[Wang and Cardie, 2014] Lu Wang and Claire Cardie. Improving agreement and disagreement identification in online discussions with a socially-tuned sentiment lexicon. *ACL 2014*, page 97, 2014.