ACQuA at Same Side Stance Classification 2019

Alexander Bondarenko[1] **Ekaterina Shirshakova** **Niklas Homann** **Matthias Hagen**

Martin-Luther-Universität Halle-Wittenberg
[1]alexander.bondarenko@informatik.uni-halle.de

Abstract

We describe the ACQuA team's participation in the "Same Side Stance Classification" shared task (are two given arguments both on the pro or con side for some topic?) that was run as part of the ArgMining 2019 workshop.

1 Introduction

In recent years, the popularity of social media and discussion platforms has lead to online pro and con argumentation on almost every topic. Still, since not all contributions in such online discussions clearly indicate their stance or polarity, automatically identifying some post's stance could help readers quickly get an overview of a discussion similar to debating portals with pro/con arguments.

In this extended abstract, we report on our participation at the "Same Side Stance Classification" shared task. The task was run as a pilot at the ArgMining 2019 workshop and stated the problem as: given two arguments, decide whether either both support or both attack some controversial topic like gay marriage—i.e., whether the two arguments are "on the same side."

Given that the available time prior to the pilot edition of the shared task was rather limited, we decided to focus our research interest simply on examining the effectiveness of simple word n-gram features and various variants of sentiment detection for same side classification. We experiment with three respective classifiers: (1) a simple rule-based method "counting" positive and negative terms, (2) a rule-based method with sentiment flipping that uses sentiment and shifter lexicons, and (3) a gradient boosting decision tree-based method using word n-gram as features.

Not too surprisingly, the evaluation results for the three classifiers show that relying on sentiment words or word n-grams alone cannot really solve stance classification. Our "best" models achieve an accuracy of 0.54 on binary-labeled balanced test sets—obviously only a very slight improvement over random guessing.

2 Related Work

Stance classification has been studied in numerous research publications proposing different features. For instance, Walker et al. (2012) analyzed 11 feature types and showed that Naïve Bayes using POS tags achieved better results than word unigrams, while HaCohen-Kerner et al. (2017) applied an SVM classifier on 18 feature types extracted from tweets (hashtags, slang and emojis, POS tags, character and word n-grams, etc.) and reported good performance for character skip n-grams. Nevertheless, word n-grams have been a very common choice in many stance classification experiments.

Also common for stance classification is the utilization of sentiment attributes. For instance, Somasundaran and Wiebe (2010) combined argumentation-based features (1- to 3-grams extracted from sentiments and argument targets) with sentiment-based features (sentiment lexicon with negative and positive words).

Comparing different classification models, Liu et al. (2016) in their evaluation showed gradient boosting decision trees to outperform SVMs for stance classification. More recently, neural approaches have been successfully applied to stance classification: Popat et al. (2019) tuned BERT with hidden state representations, and Durmus et al. (2019) used BERT fine-tuned with path information extracted from argument trees for 741 topics from kialo.com.

Given the limited time prior to the shared task, we simply wanted to test word n-grams (gradient boosting tree-based classifier) and sentiment features (rule-based classifiers) as common feature types for stance classification.

3 Task and Data

The "Same Side Stance Classification" shared task has two experimental settings: *within-topic* (argumentative topics for training and test are the same) and *cross-topic* (argumentative topics for training and test are different).

The provided data are argumentative topics and corresponding pairs of arguments collected from the debating portals idebate.org, debatepedia.org, debatewise.org, and debate.org. The data is split into training (within-topic: 63,903 argument pairs for the two topics abortion and gay marriage, cross-topic: 61,048 argument pairs for the topic abortion) and test sets (within-topic: 31,475 argument pairs for the two topics abortion and gay marriage, cross-topic: 6,163 argument pairs for the topic gay marriage). We randomly split the provided training sets into local training, validation and test sets (80:10:10).

4 ACQuA Runs

Our three runs[1] are based (1) on a rule-based classifier, (2) on a rule-based classifier with sentiment flipping, and (3) on gradient boosting decision trees.

Rule-based classification Argument stances can either support or attack some argumentative topic. In other words, they can convey a positive or a negative "sentiment" towards the topic. Since the shared task is topic-agnostic (i.e., there is no need to distinguish topic-specific argumentation vocabulary), our first run only tries to identify whether a pair of arguments expresses the same sentiment. So far, a plethora of approaches have been proposed to classify sentiment of opinions as positive or negative (or neutral), but given the time constraints of task participation we decided to investigate whether sentiment signals in the simplest form of lexicon-based counts of positive or negative terms can contribute to same side classification.

Employing the Hu and Liu (2004)'s sentiment lexicon, we use sentiment marker keyword lists for sentiment detection (e.g., good vs. bad). Depending on whether the positive or the negative markers have a higher total count, the rule-based classifier assigns the respective label to the argument—note that sentiment flipping terms (e.g., *not* bad) are not part of our first run.

[1]Code available at: https://github.com/webis-de/argmining19-acqua-same-side/

In case that the counts of positive and negative markers are equal or if an argument does not contain any marker, a random label is assigned. This is the case for about 25% of the provided within-topic and about 20% of the provided cross-topic training pairs (12% of within and about 19% for cross-topic test pairs).

Finally, if the counter-based sentiments of an argument pair agree, they are classified as "same side."

Rule-based classification with sentiment flipping We re-implemented a sentiment classifier which is one step in a three-step approach to classify a single claim's stance as pro or con with respect to some controversial topic proposed by Bar-Haim et al. (2017). A complete approach combines argument target identification with sentiment detection and consistency/contrastiveness classification. In a semester-long student project, we re-implemented parts of this approach and verified that it produces results similar to the originally reported performances.

In the setting of the "Same Side Stance Classification" shared task, we applied only the sentiment classifier, which follows the approach by Ding et al. (2008) and uses the sentiment words counts matched with the lexicon of Hu and Liu (2004)(the same that is used in our first approach) and the shifter lexicon of Polanyi and Zaenen (2006) (sentiment shifters flip the polarity of sentiment words). We could not directly apply the target identifier and the contrast classifier due to differences in semantic structures of the IBM and Same Side datasets.

In case that the counts of positive and negative sentiments are equal or if an argument does not contain any sentiments, a label that arguments are on the same side is assigned (this reflects the majority label in the IBM dataset). This is the case for about 4% of the provided within-topic and about 0.3% of the provided cross-topic pairs in the official test set.

Gradient boosting decision tree In our third run, we use the fast gradient boosting framework LightGBM (Ke et al., 2017) that employs tree-based learning algorithms. LightGBM is often used for text classification tasks, even in one of the winning approaches in the Kaggle competition on identifying duplicate Quora questions (Iyer et al., 2017). We use token frequencies and tf-idf-weighted bags of 1-, 2-, 3-, 1–2-, and 1–3-gram

Table 1: Classification accuracy on our local test set.

Model	within-topic	cross-topic
Rule-based	0.51	0.51
Rule-based with flipping	0.50	0.50
LightGBM	0.54	0.52
Informed guessing	0.50	0.50

Table 2: Classification accuracy on the official test set.

Model	within-topic	cross-topic
Rule-based	0.54	0.50
Rule-based with flipping	0.50	0.50
LightGBM	0.51	0.50
Informed guessing	0.50	0.50

lemmas as features (often used in text classification tasks).

As LightGBM returns a confidence for predictions, we run preliminary experiments with different thresholds on our local training and validation sets to select the best performing parameters. The following features and thresholds achieved the highest accuracy in these pilot experiments: tf-idf-weighted unigram lemmas and a confidence threshold of 0.520 for the within-topic setup and of 0.501 for the cross-topic setup.

5 Experiments and Results

We use our local training, validation, and test sets (80/10/10) to train, validate, and test the LightGBM-based classifier and only test the two rule-based classifiers (they do not have training step) locally (classification accuracies on the local test set given in Table 1). The simple rule-based and LightGBM approaches perform only very slightly better than a random guessing informed about the balanced data (50:50 same / different side). One possible reason for the rule-based classifier without flipping probably is that about 25% of the cases were randomly decided due to ties in the numbers of positive/negative terms. Surprisingly, considering sentiment flipping only worsened the performance. In case of the LightGBM approach, probably simple word n-gram lemmas are still not sufficient as features for a stance classification decision tree.

Even though our approaches performed very poorly on the local data, we submitted all three approaches with their best parameter settings as runs for the shared task. To this end, the LightGBM-based approach was trained on the full official training set.

The accuracies for all our three runs as reported by the task organizers are shown in Table 2. Not too surprisingly, also on the official test set, the performance of the rule-based approaches and of the LightGBM-based approach does not really improve upon an informed random guessing (50:50 label balance). Note that the rules' without flipping slightly better performance on the official test set compared to our local test set might be due to the fewer random decisions in case of ties for the numbers of positive/negative dictionary words (12% vs. 25%).

6 Conclusion

We have submitted three approaches to the shared task on same side stance classification (i.e., deciding whether two arguments are "on the same side" for a given topic): (1) a simple rule-based sentiment-oriented approach, (2) a rule-based sentiment classifier with flipping, and (3) gradient boosted decision trees with tf-idf-weighted unigram lemmas as features.

All our runs do not really improve upon an informed random guessing. Sentiment in the simplistic form of our rule-based models does not seem to help too much in same side classification.

A proper adaptation of the complete IBM Research's stance classifier to the Same Side classification task and training classifiers over word embeddings including deployment of the neural classifiers are interesting directions for future research.

Acknowledgments

This work has been partially supported by the Deutsche Forschungsgemeinschaft (DFG) within the project "Answering Comparative Questions with Arguments (ACQuA)" (grant HA 5851/2-1) that is part of the Priority Program "Robust Argumentation Machines (RATIO)" (SPP-1999).

References

Roy Bar-Haim, Indrajit Bhattacharya, Francesco Dinuzzo, Amrita Saha, and Noam Slonim. 2017. Stance Classification of Context-Dependent Claim. In *Proceedings of ACL 2017*, pages 251–261.

Xiaowen Ding, Bing Liu, and Philip S. Yu. 2008. A Holistic Lexicon-Based Approach to Opinion Mining. In *Proceedings of WSDM 2008*, pages 231–240.

Esin Durmus, Faisal Ladhak, and Claire Cardie. 2019. Determining Relative Argument Specificity and Stance for Complex Argumentative Structures. In *Proceedings of ACL 2019*, pages 4630–4641.

Yaakov HaCohen-Kerner, Ziv Ido, and Ronen Ya'akobov. 2017. Stance Classification of Tweets using Skip Char Ngrams. In *Proceedings of ECML PKDD 2017*, pages 266–278.

Minqing Hu and Bing Liu. 2004. Mining and Summarizing Customer Reviews. In *Proceedings of SIGKDD 2004*, pages 168–177.

Shankar Iyer, Nikhil Dandekar, and Kornèl Csernai. 2017. First Quora Dataset Release: Question Pairs.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In *Proceedings of NIPS 2017*, pages 3146–3154.

Can Liu, Wen Li, Bradford Demarest, Yue Chen, Sara Couture, Daniel Dakota, Nikita Haduong, Noah Kaufman, Andrew Lamont, Manan Pancholi, Kenneth Steimel, and Sandra Kübler. 2016. IUCL at SemEval-2016 Task 6: An Ensemble Model for Stance Detection in Twitter. In *Proceedings of SemEval-2016*, pages 394–400.

Livia Polanyi and Annie Zaenen. 2006. Contextual Valence Shifters. In *Computing Attitude and Affect in Text: Theory and Applications*, pages 1–10. Springer.

Kashyap Popat, Subhabrata Mukherjee, Andrew Yates, and Gerhard Weikum. 2019. STANCY: Stance Classification Based on Consistency Cues. In *Proceedings of EMNLP-IJCNLP 2019*, pages 6412–6417.

Swapna Somasundaran and Janyce Wiebe. 2010. Recognizing Stances in Ideological Online Debates. In *Proceedings of the Workshop CAAGET at NAACL HLT 2010*, pages 116–124.

Marilyn A. Walker, Pranav Anand, Rob Abbott, Jean E. Fox Tree, Craig Martelly, and Joseph King. 2012. That is Your Evidence?: Classifying Stance in Online Political Debate. *Decision Support Systems*, 53(4):719–729.