Analysing the Submissions to the Same Side Stance Classification Task

Yamen Ajjour
Bauhaus-Universität Weimar, Germany
yamen.ajjour@uni-weimar.de

Khalid Al-Khatib
Leipzig University, Germany
khalid.alkhatib@uni-leipzig.de

Abstract

This paper presents an analysis of the submissions to the first shared task on same-side stance classification. The analysis draws attention to the potential of combining the submissions in ensemble models, demonstrates the cases where the top-performed submissions succeed to resolve the same-side stance and where they did not, and puts forward some suggestions to enhance the datasets used in the shared task.

1 Introduction

The recently proposed task of the *same-side stance classification* aims at identifying whether two arguments share the same or different stance toward a given topic. Approaching this task, the first shared-task competition [1] was introduced in the second symposium of the RATIO priority program [2] and conducted in the ArgMining workshop at ACL 2019 [3]. In this shared task, two sets of arguments that belong to the topics of *abortion* and *gay marriage* were sampled from the args.me corpus (Ajjour et al., 2019) and prepared for two experimental settings: cross-topics and within-topic. Eleven different systems were submitted to this shared task. These systems employed several supervised classifiers with various features, achieving effectiveness that ranges between 0.5 and 0.77 in terms of accuracy.

The paper at hand illustrates diverse insights for the same-side classification based on analysing the systems submitted to the shared task. In particular, we examine the effectiveness of aggregating the submitted classifiers by combining them with two ensemble models (*majority* and *oracle*). The two models were evaluated against the two experimental settings.

In addition to analyzing the ensemble models, we scrutinize the data cases which most of the classifiers tackle successfully (i.e., easy cases), and the cases in which most of the classifiers fail (i.e., hard cases). Also, we conduct a manual inspection analysis of the task data, bringing to light its limitations and proposing several suggestions to enhance it.

Our experiments show that while the *majority* ensemble is comparable to the best systems, the *oracle* ensemble achieves the optimal effectiveness. This shows that almost all the instances in the test dataset were classified correctly by at least one submitted system. The inability of the majority system to outperform the submitted classifiers shows the dominance of the top two systems (Trier University and Leipzig University). Overall, the results show the potential of using ensemble models to tackle the same-side stance classification task.

Regarding the case inspection, we discover diverse easy cases for the classifiers including when the stance towards the topic is stated explicitly using a linguistic indicator, when an argument questions certain statements in the other argument of

[1] https://events.webis.de/sameside-19/
[2] http://ratio.sc.cit-ec.uni-bielefeld.de/events/yearly-symposium-may-2019/
[3] argmining19.webis.de

	Within-Topic			Cross-Topics		
Team	**Pre**	**Rec**	**Acc**	**Pre**	**Rec**	**Acc**
Trier University	**0.85**	0.66	**0.77**	**0.73**	**0.72**	**0.73**
Leipzig University	0.79	**0.73**	**0.77**	0.72	**0.72**	0.72
IBM Research	0.69	0.59	0.66	0.62	0.49	0.60
TU Darmstadt	0.68	0.52	0.64	0.64	0.59	0.63
Düsseldorf University	0.70	0.33	0.60	0.72	0.53	0.66

Table 1: The results of the submissions which achieved more than 0.6 accuracy in the within-topic experiments and the cross-topics experiment in terms of precision (Pre), recall (Rec), and accuracy (Acc).

the pair, and when the two arguments embody contradicting statements. For the hard cases, we notice that the classifiers fail to predict the correct stance when the knowledge about the discussed topic is insufficient to resolve the stance as well as when the two arguments have partial agreement/disagreement.

Lastly, for improving the shared task datasets, we observe some problems in the data such as the treatment of debate meta-information as arguments. Based on our investigation of web resources, we propose a suggestion to sample higher quality data for the task.

2 Submission Ensembles

In this section, we first report on the results of the individual classifiers which were submitted to the shared task. Then, we present the two ensembles (*oracle* and *majority*), comparing their effectiveness to those of the individual classifiers.

2.1 Classifiers Effectiveness

To exclude potential noise that may be introduced by ineffective classifiers, we consider here only those classifiers which achieved an accuracy higher than 0.6 in both cross-topics and within-topic experiments.

Table 1 shows the results of the classifiers which satisfied our quality constraint. This constraint applies to five classifiers out of the eleven submitted classifiers.

2.2 Combined Results: Ensembles

The ensemble models, used to aggregate the classifiers, combine the predictions of the submitted classifiers in a *majority* as well as an *oracle* ensemble. Both ensembles utilize the predictions of the most effective submitted classifiers. The majority ensemble predicts the stance label of an argument pair using the majority vote of the classifiers' predictions, while the oracle ensemble uses the ground-truth labels to pick the classifier with the correct predicted label if one exists.

	Within-Topic			Cross-Topics		
Ensemble	**Pre**	**Rec**	**Acc**	**Pre**	**Rec**	**Acc**
Oracle	0.99	1	1	1	0.99	1
Majority	0.82	0.64	0.75	0.75	0.6	0.7

Table 2: The results of the ensemble classifiers oracle and majority for the within-topic experiments and the cross-topics experiment in terms of precision (Pre), recall (Rec), and accuracy (Acc)

.

Table 2 shows the results of the *oracle* and *majority* ensembles in the cross-topics and within-topic experiments. The *oracle* ensemble reaches an accuracy of 1 in both experiments. This shows that combining several classifiers for tackling the same-side classification task is a promising direction to pursue. The results also show that almost all instances in the test dataset were classified correctly by at least one system. In comparison to the top classifier, the *majority* ensemble achieves subpar accuracy in both experiments. Still, it achieves a precision of 0.75 in the cross-topic experiment, which is 0.02 points higher than the top classifier (Trier University). Besides, the majority ensemble achieves higher precision than the second top classifier (Leipzig University). The inability of the majority ensemble to enhance over the best systems in overall terms signals the superiority of the top systems (Trier University and Leipzig University) over the other three systems. However, since all instances in the test dataset were classified correctly by the different systems, it seems that the different systems learned different patterns about the task.

3 Case Analysis

In this section, we present the outcomes of manually analyzing the predictions of the eleven systems submitted to the shared task. We examine the argument pairs which are classified correctly (or wrongly) by most of the systems. A careful review of these pairs reveals some easy and hard cases for the same-side stance classification. In the following, we discuss these cases in detail.

3.1 Easy Cases

In total, we found 1234 pairs in which all the submitted systems classified correctly, 1215 in the cross-topics experiment and 19 in the within-topic. From these pairs, we determined four cases where classifying the same-side stance is doable computationally (i.e., easy cases):

1. The stance towards the same topic is expressed explicitly in the two arguments:

   Argument 1. ...because i don't believe in *gay marriage* ...

   Argument 2. ...i want to first off point out that i am against *gay marriage* personally ...

2. The two arguments include contradicting statements:

   Argument 1. ...*marriage* is not a recognition of love and compassion ...

   Argument 2. *marriage* is about love. ...

3. An argument questions a certain statement in the other argument:

   Argument 1. people should be allowed to make their own choices in life with out having their human rights taken away.

   Argument 2. i would like to know how people making their own choices has their rights taken away in the first place. give me something to argue about!

4. An argument quotes a certain statement in the other argument:

   Argument 1. i also gave references stating that in the bible homosexuality isn't even accepted.

   Argument 2. *"i also gave references stating that in the bible homosexuality isn't even accepted"* oops - sorry - the bible isn't admissible as a source of law in the us.

3.2 Hard Cases

In the test dataset, 126 argument pairs were difficult to be classified by the systems (125 in the cross-topics experiment). Two cases were noticeable in these pairs:

1. Further knowledge about the discussed topic is needed to resolve the stance:

   Argument 1. *gay marriage* violates religious freedoms

   Argument 2. *gay marriage* is a negligible change to institution of marriage

2. The two arguments agree on one aspect related to the topic but disagree on other aspects:

   Argument 1. marriage is a euphemism for using the government to enforce a relationship. there's no problem with gays getting married, but they shouldn't marry with government involvement.

   Argument 2. i say we let the gays get married. it's not like it affects anyone but them anyway.

4 Data Quality

The shared task datasets are derived from args.me corpus (Wachsmuth et al., 2017b). This corpus incorporates five different debate platforms: four comprise arguments in a monological form, while one embraces arguments within dialogues (aka debates). Because the latter is the largest platform that contributes the most to the args.me corpus with more than 182,198 arguments (%63), it largely dominates the shared tasks datasets.

Deriving arguments from dialogues, however, requires extensive preprocessing, including removing meta-dialogue and meta-users information, de-contextualizing arguments, and filtering low-quality texts that contain abusive language or spams.

This preprocessing step was not performed for the shared task datasets, which lead to several invalid argument instances. Overall, we found two main problematic cases:

1. The argument addresses solely a debate meta-information:

> **Argument .** this round is for acceptance only. the rest will be for argumentation.

> **Argument .** my opponent had forfeited the round, so my arguments stand unchallenged.

2. The argument contains ad hominom attack:

> **Argument .** like i said i didnt copy crap! and if you are going to acusse me for something i didn't do, then i wish to never have another debate with you again.

Giving that these cases frequently occur in the shared task datasets, we suggest the following improvements:

- Using only monological sources of arguments, as dialogues need the preprocessing step we mentioned above.
- Conducting manual annotation or validation of the argument pairs, especially for those which are put in the test datasets.

5 Conclusion

Analysing the output of shared tasks is key for learning lessons and prompting future development. This paper addresses the new shared task of same-side stance classification, presenting an analysis of its submissions and data. In particular, we have found that ensemble models have the potential for increasing the effectiveness of tackling the task. We also have observed that the missing knowledge of arguments and the possibility of partial agreement/disagreement between them are the main challenges of the task.

References

Yamen Ajjour, Henning Wachsmuth, Johannes Kiesel, Martin Potthast, Matthias Hagen, and Benno Stein. 2019. Data Acquisition for Argument Search: The args.me corpus. In *42nd German Conference on Artificial Intelligence (KI 2019)*. Springer.