Overview of the 2019 ALTA Shared Task: Sarcasm Target Identification

Diego Mollá
Department of Computing
Macquarie University
diego.molla-aliod@mq.edu.au

Aditya Joshi
CSIRO Data61
aditya.joshi@csiro.au

Abstract

We present an overview of the 2019 ALTA shared task. This is the 10th of the series of shared tasks organised by ALTA since 2010. The task was to detect the target of sarcastic comments posted on social media. We introduce the task, describe the data and present the results of baselines and participants. This year's shared task was particularly challenging and no participating systems improved the results of our baseline.

1 Introduction

Sarcasm is a form of verbal irony that is intended to express contempt or ridicule. Sarcastic text has been understood to be a challenge to sentiment analysis because a sarcastic text may appear to be positive on the surface but is intended to be negative. Empirical results also show that sarcastic text is detrimental to sentiment analysis (Maynard and Greenwood, 2014). Applications where sentiment understanding is important are also impacted by sarcastic text. These applications include dialogue systems where the correct prediction of sentiment is important to generate appropriate responses. Towards this, computational sarcasm has gained interest in the research community.

Sentiment in a text can be understood to be composed of valence (positive/negative) and the target (Liu, 2012). The connection between sarcasm detection and sentiment analysis is the target. Sarcastic text bears a target of ridicule. It is this target that receives negative sentiment in the sarcastic sentence.

The goal of the 2019 ALTA Shared Task is the automatic detection of sarcasm targets. Section 2 describes the general aims of the ALTA shared tasks, and the specific aim of the 2019 shared task. Section 3 briefly presents related work. Section 4 describes the data. Section 5 shows the evaluation results. Section 6 presents the results, and Section 7 concludes this paper.

2 The 2019 ALTA Shared Task

The 2019 ALTA Shared Task is the 10th of the shared tasks organised by the Australasian Language Technology Association (ALTA). Like the previous shared tasks, it targets university students with programming experience, but it is also open to graduates and professionals. The general objective of these shared tasks is to introduce interested people to the sort of problems that are the subject of active research in a field of natural language processing. Depending on the availability of data, the tasks have ranged from classic but challenging tasks to tasks linked to very hot topics of research.

There are no limitations on the size of the teams or the means that they may use to solve the problem. We provide training data but participants are free to use additional data and resources. The only constraint in the approach is that the processing must be fully automatic — there should be no human intervention.

As in past ALTA shared tasks, there are two categories: a student category and an open category.

- All the members of teams from the **student category** must be university students. The teams cannot have members that are full-time employed or that have completed a PhD.

- Any other teams fall into the **open category**.

The prize is awarded to the team that performs best on the private test set — a subset of the evaluation data for which participant scores are only revealed at the end of the evaluation period (see Section 5). The organisers reserve the right not to award the prize if no teams obtain better results than those of the published baselines.

Given a sarcastic text, the task of the 2019 ALTA shared task is to identify the set of words which are the target of sarcasm. The words are to be returned as a list with all words in lowercase, where all duplicates have been removed. If such set of words is not found, the system should return a fall-back label ”OUTSIDE”. Table 1 shows examples of sarcastic comments and the annotated targets. The assumption in each of the samples used in the shared task is that they are sarcastic.

3 Related Work

Sarcasm has been understood as a challenge for sentiment analysis (Pang et al., 2008). Over the past years, automatic detection of sarcasm gained interest. Several approaches have been reported for automatic detection of sarcasm in text, spanning rule-based approaches to deep neural architectures (Joshi et al., 2017).

Since sarcasm is a peculiar form of sentiment expression, the target of a sarcastic text bears implications on attribution of the negative sentiment to the appropriate target. For example, for an aspect-based sentiment analysis system, the sarcasm target will be the aspect towards which a negative sentiment will be assigned. Two prior papers report approaches for sarcasm target identification.

The problem of sarcasm target identification was introduced in Joshi et al. (2018). They present three kinds of methods: (a) rule-based which use heuristics to determine sarcasm targets, (b) learning-based which use a sequence labelling algorithm trained on a dataset labelled with sarcasm targets, and (c) a hybrid of the two where output of the two systems is combined to make the final predictions.

More recently, Patro et al. (2019) present a deep learning-based architecture for sarcasm target identification. The semantic representation of each word is captured in terms of its context window using a bidirectional LSTM. This semantic representation is then concatenated with features based on LIWC, NER, empathy and POS tags, to learn a classifier. They show an improvement over the prior work.

4 Data

The data used in the 2019 ALTA Shared Task consists of 950 training samples and 544 test samples. A count of the words appearing in the targets of the training data (Figure 1) reveals that a large percentage of the data is labelled as OUTSIDE, and many of the remaining words are personal and possessive pronouns, including first person “I”, “we”, “my”. This observation led us define a baseline that focus on the presence of pronouns — see Section 6 for details of the baseline.

5 Evaluation

As in previous ALTA shared tasks, the 2019 shared task was managed and evaluated using Kaggle in Class, with the name “ALTA 2019 Challenge”.[1] This allowed the participants to submit runs prior to the submission baseline for immediate feedback and compare submissions in a public leaderboard.

The test data was split into a public and a private partition. Submissions by participants were evaluated on the entire test data but only the results of the public partition were shown in the public leaderboard. Only the shared task organisers had access to the results of the private leaderboard, and these results were used for the final ranking after the submission deadline.

Each participating team was allowed to submit up to two (2) runs per day. By limiting the number of runs per day, and by not disclosing the results of the private partition, the risks of overfitting to the private test results were controlled.

The evaluation metric was the mean of the F1 score over the test samples (Formula 1),

$$\begin{array}{rcl} F1 & = & 2\frac{p\times r}{p+r} \\ p & = & \frac{tp}{tp+fp} \\ r & = & \frac{tp}{tp+fn} \end{array} \quad (1)$$

where the true positives (tp) in a sample were the set of target words correctly identified by the system, the false positives (fp) were the set of words incorrectly identified as target, and the false negatives (fn) were the set of words from the target that were not identified by the system.

The mean F-Score is equivalent to the mean of the Sørensen-Dice coefficient (Formula 2),

$$D(A,B) = 2\frac{|A\cap B|}{|A|+|B|} \quad (2)$$

where A represents the set of words of the target, and B represents the set of words of the prediction.

[1] https://inclass.kaggle.com/c/alta-2019-challenge

Comment	Target
Your shirt reminds me of my 10-year-old	your shirt
This is the best film ever!	film
Oh, and I suppose the apple ate the cheese	OUTSIDE

Table 1: Examples of sarcastic comments and their targets.

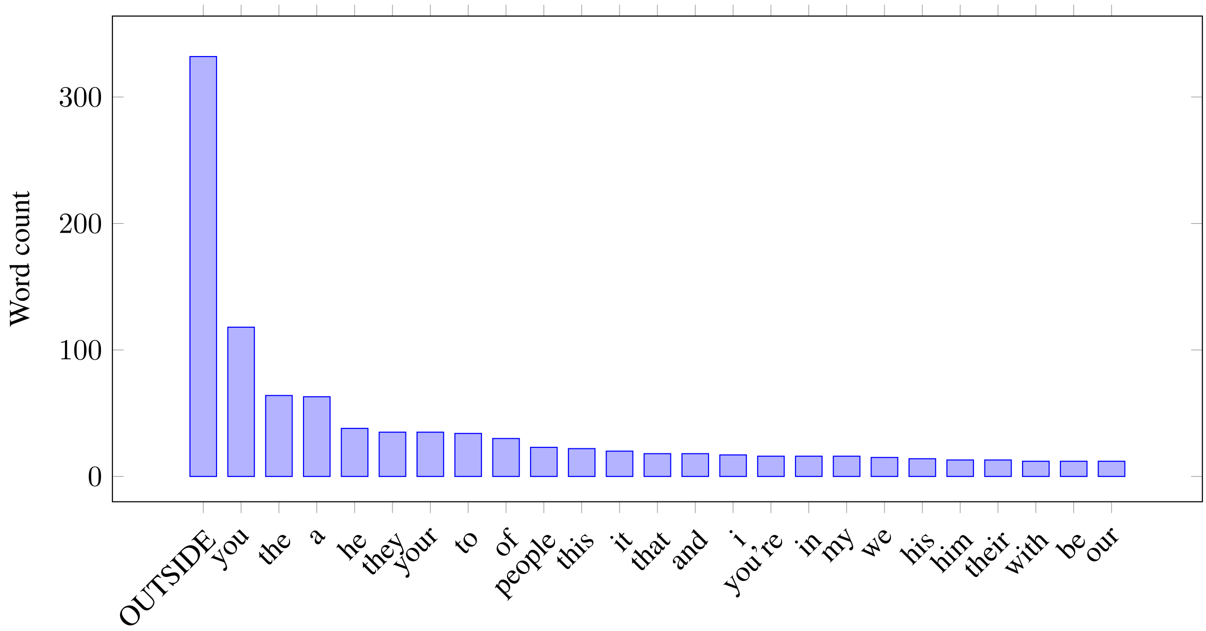

Figure 1: Most frequent words appearing in the targets of the training data.

Name	Category	Leaderboard Public	 Private
OUTSIDE	*Baseline*	*0.36764*	***0.34926***
Powers	Student	**0.38624**	0.33311
Orangutan	Student	0.37150	0.29218
Pronouns	*Baseline*	*0.20933*	*0.22539*

Table 2: Public and private leaderboards based on runs selected for the final ranking (by default these were the runs with highest score in the public leaderboard). The figures indicate the mean F1 score.

Name	Category	Leaderboard Public	 Private
OUTSIDE	*Baseline*	***0.36764***	*0.34926*
Powers	Student	0.34731	0.34490
Orangutan	Student	0.33242	**0.37802**
Pronouns	*Baseline*	*0.20933*	*0.22539*

Table 3: Public and private leaderboards based on runs with best scores in private leaderboard. The figures indicate the mean F1 score.

6 Results

Two baselines were made available to the participants as a Kaggle notebook.[2] The first baseline simply returned the word OUTSIDE, meaning that in all cases the target was predicted as not explicitly mentioned in the text. This baseline proved particularly hard to beat, as discussed below.

The second baseline is based on the observation that many of the target words are pronouns (Figure 1). Thus, the baseline returns all personal and possessive pronouns, and if no such pronouns are found, it returns OUTSIDE.

In total 16 teams registered for the competition — 14 in the student category and 2 in the open category. Of these, only 5 teams submitted runs, and only 2 submitted valid runs with results different from the baselines. Table 2 shows the results of the public and private leaderboard for the baselines and the 2 teams.

As Table 2 shows, no teams outperformed the OUTSIDE baseline in the private partition. A team was allowed to submit up to two runs per day, and the team received immediate feedback of the score of the public leaderboard. By default, the final submission was the one with the highest score in the public leaderboard, and the team had the option to override the default and select a different run. We observed that, even though none of the selected runs outperformed the baseline in the private leaderboard, some runs with lower scores in the public leaderboard did outperform the baseline in the private leaderboard. Table 3 shows the results of the best runs in the private leaderboard and their scores in the public leaderboard. The runs of Table 2, however, were not considered for the final ranking.

It is possible that the existence of the OUTSIDE label made the task particularly challenging. We therefore also conducted an alternative evaluation (not used for the final ranking) where we removed all samples labelled as OUTSIDE by either the annotators or the system (Table 4). The data set for this evaluation was the entire test data set combining the public and private partitions.[3] The table also includes the results of the pronoun baseline evaluated on the same data. None of the systems beat the pronoun baseline on the same test data.

The results of Table 4 use different data for each system and therefore they cannot be used for comparing the systems. Also we should note that the systems were designed assuming that some of the data would be labelled as OUTSIDE, so the results are probably not indicative of the quality of the systems.

7 Conclusions

The aim of the 2019 ALTA shared task was to detect the target of sarcastic comments. As in previous years, the task was managed as a Kaggle-in-Class competition. This year the task proved particularly challenging and none of the selected runs obtained better results than the baselines in the private leaderboard and therefore no prizes were given. The challenge will remain open in Kaggle in Class and new submissions are welcome.

Acknowledgments

The authors would like to thank Pranav Goel, University of Maryland, for his assistance with obtaining the dataset.

[2] https://inclass.kaggle.com/dmollaaliod/baselines-for-sarcasm-target-identification

[3] The reason behind using the combine public and private partitions was that the information about what samples belonged to each partition was not available in the Kaggle in Class platform.

Name	Category	Mean F1	*Mean F1 of Pronoun Baseline*	Test Size
Powers	Student	0.37931	*0.38152*	170
Orangutan	Student	0.35469	*0.31534*	105

Table 4: Evaluation on the test data after removing entries labelled as OUTSIDE by annotators and systems. The figures indicate the highest mean F1 score of each of the participant's submissions, which could be a different submission from the systems of Tables 2 and 3.

References

Aditya Joshi, Pushpak Bhattacharyya, and Mark J. Carman. 2017. Automatic sarcasm detection: A survey. *ACM Computing Surveys*, 50(5):73:1–73:22.

Aditya Joshi, Pranav Goel, Pushpak Bhattacharyya, and Mark J. Carman. 2018. Sarcasm target identification: Dataset and an introductory approach. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Bing Liu. 2012. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167.

DG Maynard and Mark A Greenwood. 2014. Who cares about sarcastic tweets? investigating the impact of sarcasm on sentiment analysis. In *LREC 2014 Proceedings*. ELRA.

Bo Pang, Lillian Lee, et al. 2008. Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, 2(1–2):1–135.

Jasabanta Patro, Srijan Bansal, and Animesh Mukherjee. 2019. A deep-learning framework to detect sarcasm targets. In *Conference on Empirical Methods in Natural Language Processing*, Hong Kong, China. Association for Computational Linguistics.