Overview of the 2020 ALTA Shared Task: Assess Human Behaviour

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

Abstract

The 2020 ALTA shared task is the 11th instance of a series of shared tasks organised by ALTA since 2010. The task is to classify texts posted in social media according to human judgements expressed in them. The data used for this task is a subset of SemEval 2018 AIT DISC, which has been annotated by domain experts for this task. In this paper we introduce the task, describe the data and present the results of participating systems.

1 Introduction

Human behaviour can be negatively or positively assessed based on a reference set of social norms. When judgement is explicitly stated in narratives, e.g., "They are hard-working and honest.", we can attempt to encounter appraisal words such as "hard-working" and "honest" used between interlocutors for advancing their judgement.

Attitude positioning plays an important role in Martin and White's (2005) Appraisal framework[1] (AF) for analysing someone's use of evaluative language to negotiate solidarity.

To the best of our knowledge, no prior work has attempted to automatically codify text using the AF judgement categories. The goal of the 2020 ALTA shared task is to develop a computational model that can identify and classify judgements expressed in textual segments. Participants are challenged to predict the judgement appraised by classifying each short-text message into one or more label candidates (or none): *normality, capacity, tenacity, veracity, propriety*.

2 The 2020 ALTA Shared Task

The 2020 ALTA Shared Task is the 11th of the shared tasks organised by the Australasian Language Technology Association (ALTA). As in 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. Details of the 2020 ALTA Shared task and past tasks can be found in the 2020 ALTA Shared Task website.[2]

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.

3 The Appraisal Framework

The Appraisal framework (AF) is concerned with the use of linguistic markers for identifying and track the ways attitudes are invoked in authored

[1] https://www.grammatics.com/appraisal/

[2] http://www.alta.asn.au/events/sharedtask2020/

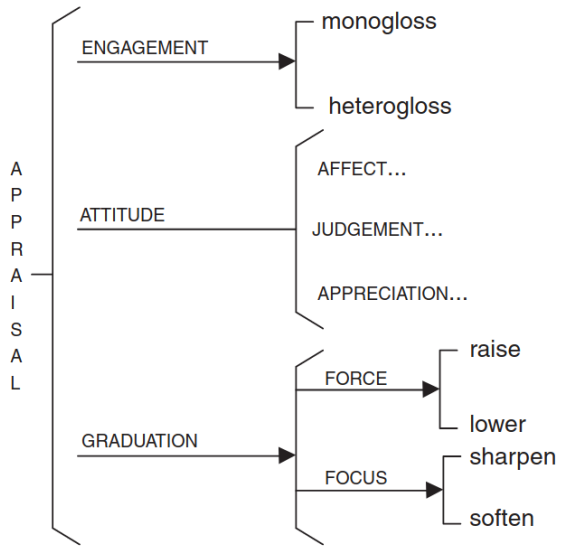

Figure 1: Overview of appraisal resources (Martin and White, 2005, p38)

text. The framework defines three subsystems for evaluative meaning making (1) ATTITUDE; (2) ENGAGEMENT; and (3) GRADUATION. Each of these are further divided in to other subsystems (Figure 1). In particular, The ATTITUDE framework is divided into three subsystems: (1) AFFECT (registering of emotions); (2) APPRECIATION (evaluations of natural and semiotic phenomena); and (3) JUDGEMENT (evaluations of people and their behaviour).

The judgement subsystem has two regions: social esteem and social sanction. The subcategories of each of these two regions form the target labels for the 2020 ALTA Shared Task. In particular:

Social esteem tends to function as admiration or criticism and can be subdivided into three subcategories:

Normality (how unusual one is): "He is old-fashioned".

Capacity (how capable one is): "Self-driven 12 year old is a maths genius".

Tenacity (how resolute one is): "They are hard-working and honest".

Social sanction functions as praise or condemnation and can be subdivided into two subcategories:

Veracity (how honest/truthful one is): "They are hard-working and honest".

Propriety (how ethical one is): "She is too arrogant to learn the error of her ways".

The judgement system is used to assess human behaviour and their position on certain social norms. Further details and examples can be found in The Appraisal Website.[3]

4 Data

The source data of the 2020 ALTA Shared Task is a subset of the SemEval 2018 AIT DISC dataset.[4] A total of 300 tweets have been manually annotated in a two-stage process. The annotation was first annotated by two linguists from two Australian universities (University of Wollongong and University of New South Wales) and then double-checked by two other linguists from the same two universities. The data were subsequently split into a training set of 200 tweets, and a test set of 100 tweets.

Each tweet was annotated with one or more (or none) of the following labels: *normality, capacity, tenacity, veracity, propriety*. Table 1 shows artificial examples of text messages and their annotations.

5 Evaluation

As in previous ALTA shared tasks, the task was managed as a Kaggle in Class competition. This year's task name was "ALTA 2020 Challenge".[5] The Kaggle-in-Class platform enabled the participants to download the data, submit their runs, and observe the results of their submissions in a leaderboard instantly.

As is common in Kaggle competitions, when a participant team submits their results, the public leaderboard shows the evaluation results of part of the test data, and the results of the remaining test data are held for the final ranking. By following the public leaderboard, a team can then gauge the performance of their system in comparison with that of other systems in the same public test set. A team can choose up to two of their runs for the final ranking. If a team chooses runs for the final ranking, the best results on these runs on the private partition of the test data will be used. If a team

[3] https://www.grammatics.com/appraisal/appraisalguide/unframed/stage2-attitude-judgement.htm

[4] https://competitions.codalab.org/competitions/17751#learn_the_details-datasets

[5] https://www.kaggle.com/c/alta-2020-challenge/

Text	Normality	Capacity	Tenacity	Veracity	Propriety
Read and try to comprehend what you have commented on.	0	1	0	0	0
Fans of adoring Dictatorships and Totalitarians.	0	0	0	0	1
Keep going like you always have done.	0	0	1	0	0
She showed her true colors.	0	0	0	0	1
He is a nasty person.	1	0	0	0	1
Corruption 101	0	0	0	1	0

Table 1: Artificial examples of texts and their annotations.

does not choose any runs, the private evaluation results of the run with the best results on the public partition will be chosen.

The systems were evaluated using the mean of the F1 score over the test samples (1),

$$\begin{aligned} F1 &:= \frac{1}{|S|}\sum_{s\in S} F_\beta(y_s,\hat{y}_s) \\ F_1(y_s,\hat{y}_s) &:= 2\frac{P(y_s,\hat{y}_s)\times R(y_s,\hat{y}_s)}{P(y_s,\hat{y}_s)+R(y_s,\hat{y}_s)} \\ P(y_s,\hat{y}_s) &:= \frac{|y_s\cap\hat{y}_s|}{|y_s|} \\ R(y_s,\hat{y}_s) &:= \frac{|y_s\cap\hat{y}_s|}{|\hat{y}_s|} \end{aligned} \quad (1)$$

where y_s is the set of predicted labels in sample s, $\hat{y}_s$) is the set of true labels in the sample, and S is the set of samples. If there were no true or no predicted labels, $F_1(y_s,\hat{y}_s) := 0$.

6 Participating Systems

In total 5 teams registered for the competitions, all of them in the student category. Of these, 3 teams submitted runs.

Team NLP-CIC experimented with logistic regression and Roberta (Aroyehun and Gelbukh, 2020). Whereas the logistic regression classifier obtained the best results in the public leaderboard, it performed much worse in the private leaderboard. In contrast, the Roberta classifier obtained consistent results in both the public and private leaderboards.

Team OrangutanV2 designed classifiers using ALBERT and transfer learning (Parameswaran et al., 2020). After observing that 22 tweets from the test set are also in the training set, they also incorporated a component that performed cosine similarity with the samples from the training data.

Team NITS experimented with ensemble approaches (Khilji et al., 2020). They obtained pretrained word embeddings and incorporated polynomial features. These features were fed to decision tree and Extreme Gradient Boosting (XGBoost) classifiers.

7 Results

Table 2 shows the results of the systems in the private leaderboard.

Team	F1	p
NLP-CIC	0.155	
OrangutanV2	0.105	0.313
NITS	0.053	0.010

Table 2: Results of the participating teams according to the private leaderboard. Column p indicates the Wilcoxon Signed Rank test between a team and the top team after removing ties.

The results indicate that this task has been particularly challenging and there is room for improvement. A possible reason for the difficulty of this task is the small number (200) of annotated samples available. Another reason for the low results is the relatively large percentage of samples with empty judgements. In particular, 60% of the test data had empty judgements. According to Formula (1), the F1 score of test samples with no annotations is 0. This means that the upper bound with this test data is 0.4.

8 Conclusions

The aim of the 2020 ALTA shared task was to predict the judgement of short texts according to Martin and White's (2005) Appraisal framework. The task proved challenging, presumably due to the small amount of annotated data and the sparse annotations in the data.

Acknowledgments

We thank the anonymous sponsor who donated the data for this shared task.

References

Segun Taofeek Aroyehun and Alexander Gelbukh. 2020. Automatically predicting judgement dimensions of human behaviour. In *Proceedings of the 18th Annual Workshop of the Australasian Language Technology Association*.

Abdullah Faiz Ur Rahman Khilji, Rituparna Khaund, and Utkarsh Sinha. 2020. Human behavior assessment using ensemble models. In *Proceedings of the 18th Annual Workshop of the Australasian Language Technology Association*.

J. Martin and P. White. 2005. *The Language of Evaluation Appraisal in English*. Palgrave Macmillan, UK.

Pradeesh Parameswaran, Andrew Trotman, Veronica Liesaputra, and David Eyers. 2020. Classifying JUDGEMENTS using transfer learning. In *Proceedings of the 18th Annual Workshop of the Australasian Language Technology Association*.