Automatically Predicting Judgement Dimensions of Human Behaviour

Segun Taofeek Aroyehun & Alexander Gelbukh
CIC, Instituto Politécnico Nacional Mexico City, Mexico

Abstract

We describe our submission to the ALTA-2020 shared task on assessing behaviour from short text, We evaluate the effectiveness of traditional machine learning and recent transformers pre-trained models. Our submission with the Roberta-large model and prediction threshold achieved first place on the private leaderboard.

Background

- Language enables us to express evaluation of people, action, event, and things
- The appraisal framework of [2] provides a detailed classification scheme for understanding how evaluation is expressed and implied in language
- Three categories of evaluative text: affect, judgement, and appreciation
- Utterances are viewed as indicating positive (“praising”) or negative (“blaming”) disposition towards some object (person, thing, action, situation, or event)
- The judgement dimensions are normality, capacity, tenacity, veracity, and propriety

Each of the dimensions represents an answer to the following corresponding questions:

- Normality: How special?
- Tenacity: How dependable?
- Capacity: How capable?
- Veracity: How honest?
- Propriety: How far beyond reproach?

Task Description

Given a short text, predict one or more judgement dimensions expressed in the given text. This is a multilabel classification problem where the labels consist of the five judgement dimensions.

Data

We used the data provided by the organizers of the ALTA-2020 shared task [3]. The training set has 198 tweets and the test set consists of 100 examples.

Label	Proportion
Normality	0.11
Capacity	0.16
Tenacity	0.11
Veracity	0.015
Propriety	0.18

Table 1: Frequency of each label in the training set as a fraction of the total number of training examples.

Models

- **NBSVM** uses the naive bayes log-count ratio of n-grams as features [4]. They are fed into a logistic regression classifier. We train a binary classifier per label.
- **Roberta-large** is an optimized BERT model trained on a larger and more diverse collection of text [5]. We fine-tune the pre-trained model on the training dataset provided for the shared task.

Acknowledgement

We thank CONACYT for the computer resources provided through the INAOE Supercomputing Laboratory's Deep Learning Platform for Language Technologies.

Contact Us

Segun: aroyehun.segun@gmail.com & https://nlp.cic.ipn.mx/segun/
Alexander: gelbukh@gelbukh.com & https://www.gelbukh.com/

Experiments

The use of neural networks has led to significant performance improvements in NLP tasks. However, neural networks require a large amount of labeled data. On the contrary, the traditional machine learning models such as NBSVM are competitive in low-data regimes [1]. We examined the effectiveness of NBSVM and a Roberta-large model for predicting dimensions of judgement expressed in short text.
Data pre-processing. We clean the text of each tweet by removing punctuation marks, digits, and repeated characters. We normalize URLs and usernames (tokens that starts with the @ symbol). Hashtags are converted to their constituent word(s) after removing the # symbol.
Classifier threshold. We set 0.2 as the decision threshold for the *capacity* label and 0.1 for the remaining labels.

Results

Method	Public leaderboard	Private leaderboard	Average
NBSVM	**0.16000**	0.00000	0.08000
NBSVM w/ prep.	0.16000	0.00000	0.08000
Roberta-large	0.11666	0.06666	0.09166
Roberta-large w/ threshold	0.14285	**0.15466**	**0.14876**

Table 2: Mean F1 score on the public and private test sets obtained on kaggle In-class.

Our best model achieved the first position on the ALTA-2020 shared task.

Conclusion

- We used NBSVM and Roberta-large to automatically predict the dimensions of human judgement
- NBSVM model did not generalize
- Roberta-large model with prediction threshold was consistent
- With the small size of the test set, we cannot conclude which model is better

References

[1] Segun Taofeek Aroyehun and Alexander Gelbukh. Aggression detection in social media: Using deep neural networks, data augmentation, and pseudo labeling. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 90–97, 2018.

[2] James R. Martin and Peter R. White. *The language of evaluation*, volume 2. Springer, 2003.

[3] Diego Mollá. Overview of the 2020 ALTA Shared Task: Assess Human Behaviour. In *Proceedings of the 18th Annual Workshop of the Australasian Language Technology Association*, 2020.

[4] Sida I. Wang and Christopher D. Manning. Baselines and bigrams: Simple, good sentiment and topic classification. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 90–94, 2012.

[5] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.