Author: Jacob Fuehne

Date: 9/27/2020

Research Paper Summary

SemEval-2018 Task 1: Affect in Tweets, *Association for Computational Linguistics*, 2018

1. Problem definition and the main ideas of the research

Emotions are central to language and thought. In understanding emotions, one often works with intensity (degree or amount of emotion), as well as identifying positiveness - negativeness, active - passive, and dominant - submissive (the valence arousal dominance model or VAD model). This paper is a summary of the methods, resources, and tools used by the 75 teams that participated in SemEval-2018 Task 1. The subtasks for this include: emotion intensity regression, emotion intensity ordinal classification, valence (sentiment) regression, valence ordinal classification, and emotion classification.

2. Significance of research study (Importance and Challenges of research problem)

This paper notes that natural language applications in commerce, public health, disaster management, and public policy are able to benefit from knowing the affectual states of people. Due to the number of teams involved in SemEval-2018 Task 1, this paper can provide some insight into the current tendencies, methods used, and bias of computational linguists. As this paper outlines the annotation methods used, it is also an important resource in annotation methods. A common challenge faced among annotators was obtaining consistent annotations, while the researcher teams faced issues in handling racial and gender bias during classification.

3. Main research questions and assumptions

This paper focuses on summarizing SemEval-2018, from its annotation methods, methods used by participating research teams, and gender and racial bias in models submitted by

teams. The paper assumes that by gathering 75 teams (200 team members) to tackle the same affective computing tasks, the authors will be able to make conclusions about the preferred methodologies and performance of teams. In turn, the authors will then be able to comment on the state of top-performing systems in affective computing in general.

4. Research Methodology

In the paper, the authors describe how they created the Affect in Tweets Dataset. Tweets are gathered from English, Arabic, and Spanish and annotated via crowdsourcing and a gold standard by the authors. Each tweet was labelled with an emotion category (ie, anger also included annoyance and rage). The authors of the paper noted the models used by teams on the various subtasks, as well as the features and resources. In their paper, they note that most of the top performing teams used both deep neural network representations (sentence embeddings) as well as features from existing sentiment and emotion lexicons.

5. Experiments

As a summary of SemEval-2018, the authors focus on describing the results of the event rather than performing their own experiments. However, the paper does present the algorithms, features, and resources used by the teams, as well as examining bias of submitted sentiment analysis systems. In analyzing models used by teams, a majority of the top performing systems used manually-engineered representations for tweets, rather than neural networks. This shows that representation learning can benefit from working with task-specific features. Using the Equity Evaluation Corpus, a collection of 8,640 English sentences chosen to tease out gender and race biases, the researchers examine bias of submissions. For gender, 75% to 86% of submissions consistently marked sentences of one gender higher than another when classifying emotion. For race, the bias was even more clear.

6. Discussion

6.1 Important aspects

- Data is made readily available to the public

  Having the data be readily available to the public helps advance the field as a whole. Many computational linguistics researchers will cite public datasets in their publications, but it is unfortunately still a common occurrence where authors will not make their own datasets/code implementations public and readily available.

6.2 Limitations of the paper

- There was limited information available regarding the gender and racial bias other than presenting them

  The authors of the paper opted to not comment on or give in depth details on the bias that they researched, instead simply presenting them. As such, this part of their paper serves no purpose other than to acknowledge the existence of a bias in this particular event.

6.3 Questions for presenter

- With regards to the clear racial bias that was shown in the evaluation of emotional intensity among participants, while it wasn't talked about in the paper, what are some possible reasons that you can think of for this bias?
- Do you think that computational linguists would be better served by being race blind in their models (ie, including proper names into a stop words list that would be ignored for emotion classification)?