CSE 574, Introduction to Machine Learning, Spring 22, Assignment 3

Sakthivadivel, Kaviarasu
Jayandran, Kadambare
Udhayasankar, Naveen

The Probables

1.2.1 (2 points) - Based on what we discussed in class about this dataset and the task of stance annotation (here, for attitudes towards vaccines), on what percent of the tweets that you annotate do you think the two annotators will agree? There is no right answer here, obviously, but provide a justification of your response.

The agreement between two annotators depends on various factors like the expertise of the annotator in the field, the personal bias of the annotator, the political standing of the annotator and the like. Since the annotators in our group have similar aspects we were expecting an agreement of about 75% while allowing some space for personal bias.

1.2.2 (3 points) - Based on what we discussed in class about this dataset, what percent of the tweets that you annotate do you think will be labeled pro-vaccine? Justify your answer.

We assumed that, for a dataset for natural language processing task, the data will be stratified with equal pro vaccine, against vaccine and neutral tweets. We were also expecting a few tweets where no stance could be taken. Going by this trend, about 40% of the tweets are expected to be pro-vaccine.

2.1 - What was your overall percent agreement, and how did this compare to your estimate from 1.2.1? Why were these similar/different?

The calculated percentage agreement was 70% while we expected a percentage agreement of 75%. Both the values are quite similar.This might be due to all the annotators having a similar expertise and perception of tweets. Had the annotators been from varying backgrounds with different levels of expertise, the percentage agreement would've been much lower. Though relatively high , this wouldn't be an acceptable value if we are looking to check its reliability for making decisions in the medical field or any other areas with higher stakes.

2.2 - Was this annotation task harder or easier than you expected? Why, and what was the hardest part?

The annotation was hard in some aspects as to agreeing between two annotators. We spent a considerable amount of time resolving conflicts between two annotators. Since we are from the same academic field, and have a similar stance in most cases we had hoped for most of our annotations to match but we had a significant level of disagreement between any two annotators. If resolving conflicts in a small dataset takes considerable human effort and time, conflict resolution for a real-world dataset will require more time.

2.3 - Does this change your perspective on how people talk about health-related content on social media? Why or why not?

The annotation task certainly has a definitive impact on how we view health related content on social media. The general consensus among the annotators when we began the annotation task was that there will be a lot of tweets supporting the vaccine as the COVID pandemic had impacted the entire planet. But as we progressed through the annotation task, we were surprised to find an equal number of tweets against the vaccine. There were also tweets which were just news and no stance can be taken from those tweets.This makes us question the reliability of the information available through social media. Given the critical nature of the information, some sort of regulation might be useful for health related content on social media.

2.4 - What did you learn from this annotation task that will change your perceptions of how NLP models are trained moving forward?

The task helped us understand the underlying complexity in deducing the sentiment with very brief statements. There were tweets that had a number of positive words though in actuality the overall standing of the user were more inclined against the idea of vaccination. We would have to take into account sentiments like sarcasm and humor. Hence, NLP models should incorporate various factors to accurately analyze the sentiment without bias or error. This also shows that training a natural language processing system with a small amount of data points is a difficult task and will not produce accurate results.

2.5 - How do you think your team's agreement (say, in terms of Krippendorf's Alpha) will compare to the other teams? Why do you think so?

Since this annotation task is academic and involves groups with annotators who have more or less the same level of expertise in the subject and most of the teams come from a similar geographical area and hence exhibit similar effects of the pandemic, we would expect our value of Krippendorf's alpha to be in alignment with the other groups. But the annotations of other groups might have a slightly higher or lower personal bias, for instance, a doctor in the family or a family member who was affected by COVID. Given these factors, we expect our Krippendorf's alpha value of 0.50 to match with about 60% of the other groups.