Gregory Wert : gaw8pa | Teresa Sun : js6sk | James Xie : jx4cc

**Twitter Sentiment Competition Write up**

Text analysis is an incredibly important and applicable form of data science. Text, being one of the primary means of communication, can provide valuable amounts insight into people’s views. Text is thus an incredibly valuable form of data that can nonetheless prove somewhat more challenging to work with. The text analysis of twitter data we performed showcases the benefits and pitfalls of such analysis. It shows the various potential benefits in working with text analysis while showing the difficulty in finding means of powerfully analyzing it.

The analysis of twitter sentiment showed how tweets can be used to estimate product interest and expectations. Accurately predicting this can provide valuable data and analysis to many interested parties. Firstly, it can tell companies how users feel about a certain product. Knowing this, a company can adjust in many ways. It can work on targeting people interested already, or it can use this to alter it course and interest new people: such as by adjusting its marketing. Similarly, they can forecast how much interest there will be in the product by seeing how people feel towards it. Views towards electric cars being positive, for example, may indicate a willingness to immediately adopt it. Conversely, regulators may use the data to understand what concerns consumers may have with a product. For example, if they predict a lot of negative sentiments, then they may want to explore those low sentiment tweets and see what the concern is. Companies and regulators may then work to address the problem before it become an issue. The value of these tweet analyses and predictions is that they give insights into how people conceive of and how people view a product: understanding this may help allay concerns and entice desires.

This problem is challenging however because deciphering people’s full views from tweets is complicated. The tweets themselves were manually marked up, so the models may be fitting towards the markers conception of positive instead of the tweeters. There may be further bias in using sentiments in tweets as a proxy for overall sentiment since. People who use twitter for example may not be an accurate reflection of the populace; this is especially true when looking at this electric car example since people taking the time to tweet about them may be a special subpopulation. Even without these biases there is room for errors and issues. In our own model, we had limited success accurately predicting the sentiments. This was likely due to overfitting the model on the training set. Text presents a variety of possible relations and finding the accurate and predictive ones can be difficult.

This problem resembles various other problems in text mining. One such example is our team member—Greg’s—capstone. In that text is being analyzed in Arabic to see whether it is associated with violence in the hopes of creating a predictive model. Similarly, there have been attempts to predict crime hotspots with twitter that resemble this. They look for certain word clusters to see how they may be associated with crimes and how that can predict it.