Statistical Analysis – AI Tweets Sentiment Analysis

Text mining and sentiment analysis of the 7-day period of Tweet texts (June 2018) about the topic of artificial intelligence (AI) found frequencies of individual meaningful words and word pairs (bigrams). Sentiment analysis for this natural language processing project showed generally more positive than negative words were used with respect to AI. Net sentiment, the result of adding frequencies of positive and negative words, for the corpus (entire pool of Tweeted message texts) is plotted for the Bing sentiment lexicon. Frequencies of the eight specific sentiments provided by the NRC sentiment lexicon were compared and contrasted visually , particularly for the most frequently occurring specific sentiments of “trust” and “fear.” The Bing and NRC lexicons allowed only analysis of binary variables which showed that positive sentiments outweighed negative sentiments.

Contextual analysis on bigrams, conducted by using negating first words of word pairs, did not add significantly to the findings of overall positive sentiments.

The AFINN sentiment lexicon provided the most opportunity for statistical analysis because AFINN scores words on the range of negative through positive, - 5 through 5, where 0 is neutral. The analysis of mean scores by Twitter user (user\_id) found few extreme average scores, either positive or negative. The distribution of the means by user revealed a bimodal distribution where the positive Tweeters greatly outnumbered the negative Tweeters. This result confirmed the net positive sentiments found in the corpus.

The short period for which the data was extracted, and the fact that there was no newsworthy event in artificial intelligence within the period, did not allow for meaningful temporal analysis (changes over time). No trends or lag times could be examined due to this data set constraint. The project was conceived under the assumption that the Twitter extraction policy would not change in the next three months which would have allowed 30-days of data pulled from the twelve month prior period. The project was designed for the Twitter data to be extracted a few days prior and a few weeks after a newsworthy AI event, like the dialogue that occurred between Elon Musk and Mark Zuckerberg about AI. This, or a similar event, would be captured in social media and ensure a sufficient randomness in population for representing the sentiments of the general tweeting public in regards to artificial intelligence.