Predictive Analysis – AI Tweets Sentiments

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) for which there were more positive than negative sentiments.

The project was undertaken to gauge the sentiments of the Tweeting public with respect to artificial intelligence. However, the data did not support creating a predicitive model that was generalizable to the general public.

From visual inspection of the initial word clouds and Twitter names, marketers and other people employed in AI are highly represented in the data set used. (The Twitter names were not shown for privacy concerns. Numeric user identifications were used for grouping texts and calculating mean sentiments. However, data for 7-days prior to extraction is available, free of charge, to any Twitter account holder.) A more representative sample should be used for a training data set to model sentiments to generalize results to other populations.

The hypothesis that a larger population of Tweeters in a 7-day period immediately following a news worthy AI event would result in less positive sentiments, even if the news were positive, could be tested. This would test the sentiments for fear of the known and unknown (combined) but also counter the indirect self-selection problem that the current data set has.

A sufficiently random and large set of tweet texts would allow for topic analysis. Using the Latent Dirichlet Allocation algorithm, topics have been modeled on the total pools of Twitter texts (corpus) for discovering what people are talking about. However, because the first criteria for selecting texts is the subject of AI, other topics would be a second layer, and due to the character limit constraints, there would need to be more Tweeters or data collected over a longer period of time than 7 days. The value of analysis of word contribution to sentiment scores suffers similarly, and would need to be aggregated by user or applied for trend analysis on the corpus over more time.

Creating a logistic regression model on a sufficiently random and large data set that would support topic analysis could be used to predict topics within the discussion of AI and their respective positive or negative sentiments. Further, the NRC lexicon could be used to speciate the sentiments by topic.

Topics within the discussion of artificial intelligence, like healthcare, business analytics, tracking, targeting and ethics could be modeled by supervised logistical regression for predicting sentiments. Unsupervised learning could assign rankings by clustering topics for sentiment analysis, and clustering user sentiment means.

This information would be helpful to particular industries that use artificial intelligence heavily to craft corporate responsiblity and outreach programs. Given transparency and ethical use of the findings, the result could be used in reputation management.