Text mining for customer comments of social media per two or three mediums (batch or stream, a telecom company preferably) was outlined and proposed as the third cap-stone project.

Traditionally, data science is polarized between variations of regression and classification and aims to make predictions. However data types are being changed in many ways, its size is (from terabytes to petabytes) getting bigger, its form switches from structured to unstructured and most importantly its volatility is getting closer (form batch to stream) to more real-time. All those changes make the analysis type changed too, as we talk more about unsupervised machine learning, rather than supervised machine learning, where the former is more event based rather than classical rule-based world of supervised system.

T-Mobile will be scrutinized in respect to Twitter and/or Better Business Bureau (bbb.org) and/or Trustpilot (truspilot.com) modalities. In total, 5K comments was planned to be analyzed. The topics to be covered will include:

* Gensim Summarization
* Sentiment with polarity/subjectivity
* Complexity with lexical diversity
* Named entity recognition
* Topic modelling (clustering) with LDA (Latent Dirichlet Model) model
* Vader sentiment
* NRC word-emotion association lexicon (EmoLex)
* Cosine similarity
* Word clouds

In the end, it is expected to support findings on Capstone # 2, in respect to:

* Listening customers (e.g. compound score) increases accuracy for churn
* Utilizing probabilities, instead of prediction itself, can be more practical for e.g. call center agent and all other stakeholders
* Doing it real time will increase the effectiveness