**Project proposal - Topics in Big data**

**Twitter data Sentiment analysis**

**Problem Statement:**

To build a model for Sentiment analysis on twitter dataset to predict whether tweet is positive or negative using BERT and logistic regression.

**Dataset**:

This is the sentiment140 dataset. It contains 1,600,000 tweets extracted using the twitter api . The tweets have been annotated (0 = negative, 4 = positive) and they can be used to detect sentiment.

The dataset contains the following 6 attributes:

1. Target: the polarity of the tweet (0 = negative, 4 = positive)
2. ids: The id of the tweet
3. date: the date and time of the tweet
4. flag: The query (lyx). If there is no query, then this value is NO\_QUERY.
5. user: the user that tweeted
6. text: the text of the tweet

It is publicly available at:

<https://www.kaggle.com/datasets/kazanova/sentiment140>

**Proposed approach**:

Proposal is to fit BERT model and logistic regression model using 70-30 training and test split. This will be done using pretrained BERT model from tensor flow in python.

**Methods to evaluate**:

For evaluating the model, I’m planning to plot the true positive rate by label, precision by label, recall by label, F-measure by label on the training and testing data and at the end, will going to compare the results of BERT model with logistic regression.