RUSSIAN TROLL TWEET IDENTIFICATION

DESCRIPTION

- Given a large dataset of Tweets from a Russian Troll Factory and equal sized set of normal tweets, train a binary classification model to identify the troll tweets

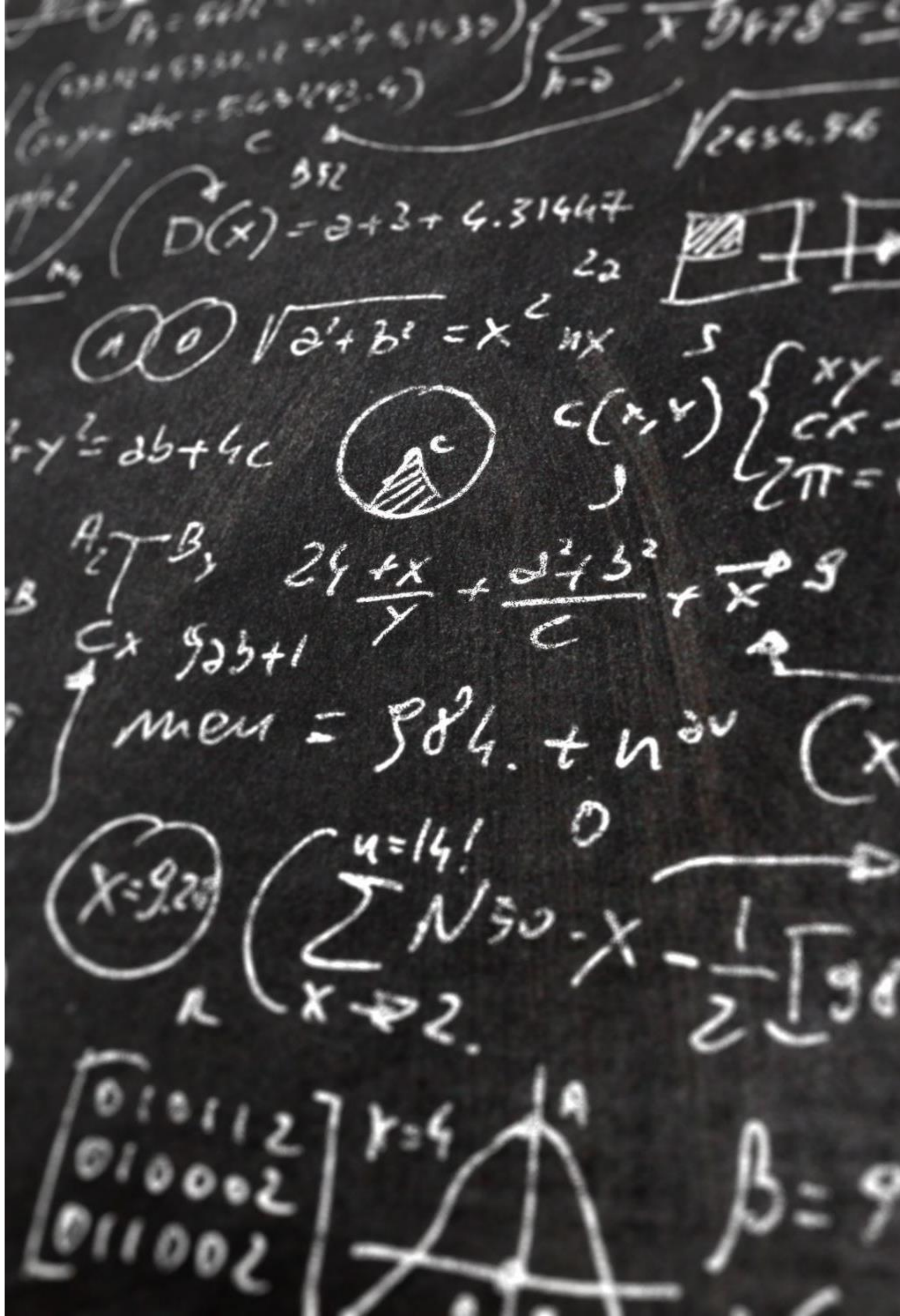
DATA SOURCES

Troll Tweet dataset: https://github.com/fivethirtyeight/russian-troll-tweets Normal Tweet dataset: https://archive.org/details/twitter_cikm_2010 The data's label will be based on its source

The ample quantity of data should be more than sufficient for supervised training of a binary classification model

TOOLS

- Numpy – for matrix and linear algebra calculations, statistics, and other mathematical computations
- Matplotlib – for graphs and other visualizations
- Scikit-learn – for converting text to numeric representations , dimensionality reduction , and testing common machine learning models
- Nltk, and spaCy – for tokenization, textual analysis , knowledge graph enhancement, and cleaning
- Gensim – for word vectorization
- Fast Sentence Embeddings – for document vectorization
- Pytorch– for custom machine learning models
- Huggingface Transformers – for transformer networks , and byte pair encoding
- Others as might prove useful such as node2vec and networkx for graph embedding and visualization



METHODOLOGY

We will design our project so that ablation analysis can be easily run by parametrizing the choices for data preprocessing, data vectorization, and model type. In this way we can measure the relative importance of the various choices that are made for these facets of modeling.

The main components that will be parametrized are:

- Data Preprocessing choices, which may include
  - Methods to limit vocabulary – frequency, Byte pair encoding
  - Stemming or lemmatization
  - Making use of POS tags
  - Markup of Named entities
  - Use of extracted phrases
  - Use of dependency trees
  - Knowledge graph enhancements using wordnet synsets

METHODOLOGY

- Data Vectorization choices, which may include
  - CountVectorizer
  - TfidfVectorizer
  - Dimensionality reduction on the above
  - Distributed word embeddings – word2vec, fasttext, glove
  - Document embeddings – doc2vec, FSE, pretrained BERT
  - Graph embeddings – node2vec

METHODOLOGY

- Classification Models
  - Perceptron
  - Logistic Regression
  - MLP
  - SVM
  - LSTM
  - CNN
  - BERT

RESULTS AND VISUALIZATIONS

Various visualizations and data analysis will drive the specific decisions made as the project progresses

We will specifically be looking for the key discriminative indicators for differentiating troll from normal tweets

T-distributed Stochastic Neighbor Embedding is a very helpful method of visualizing how classes cluster by projecting high dimensional vectors down to 2 dimensions

TIMELINE

- Week 9 – Data parsing, cleaning, and exploration
- Week 10 – Testing vectorization techniques , t-SNE visualization of class overlap
- Week 11 – Testing different models
- Week 12 – Analysis and refinement of the models, ablation studies
- Week 13 – Complete report, record presentation

RESPONSIBILITIES

- We will split the responsibilities equally across the major components of the project; data preprocessing, data analysis, data vectorization/feature extraction, modeling, analysis for model refinement, and reporting