THE STATE OF INFODEMIC ON TWITTER

BRANDON ATTAI, KELTEN FALEZ , TAHSIN CHOWDHURY

AGENDA

- Problem Overview – Classifying COVID-19 tweets as factual/false
- Extension of Original Paper Research
- Research Questions
  - Answered from
    - Dataset analysis
    - Machine learning models
- Machine learning development stage planning

PROBLEM OVERVIEW – WHAT IS BEING SOLVED?

- Identifying misinformation about covid using text(NLP) and tweet metadata
  - Classify tweet as factual/false

PROBLEM OVERVIEW – IMPORTANCE

Why is it important to solve?

- People use Twitter a news source
- False information regarding COVID-19 puts others at risk

EXTENSION OF ORIGINAL PAPER – DATA EXTENSION

Data Extension

- Add an additional 1000 samples
  - Restrict tweets from Canada

What was done in the original paper?

- 30,000 tweets obtained from multiple datasets
  - 50/50 split of factual/misinformation

EXTENSION OF ORIGINAL PAPER – EXTENDING ML

How are you going to extend the paper ML work?

- Tune hyperparameters via Gridsearch
- Feature engineering
  - Dimensionality reduction
- Additional classification models
  - Naïve Bayes classifier

RESEARCH QUESTIONS – ANALYZING DATA

Research questions that we would like to answer by analyzing the data

- Analyzing the Text:
  - How do the following in tweets with misinformation compare with those in tweets with factual information:
    - Polarity
    - Average word length
    - Length
    - Use of capital letters and punctuation
    - SMOG Index
    - The Automated Readability Index (ARI)
    - The Flesch–Kincaid ease

RESEARCH QUESTIONS ANALYZING DATA CONT'D.

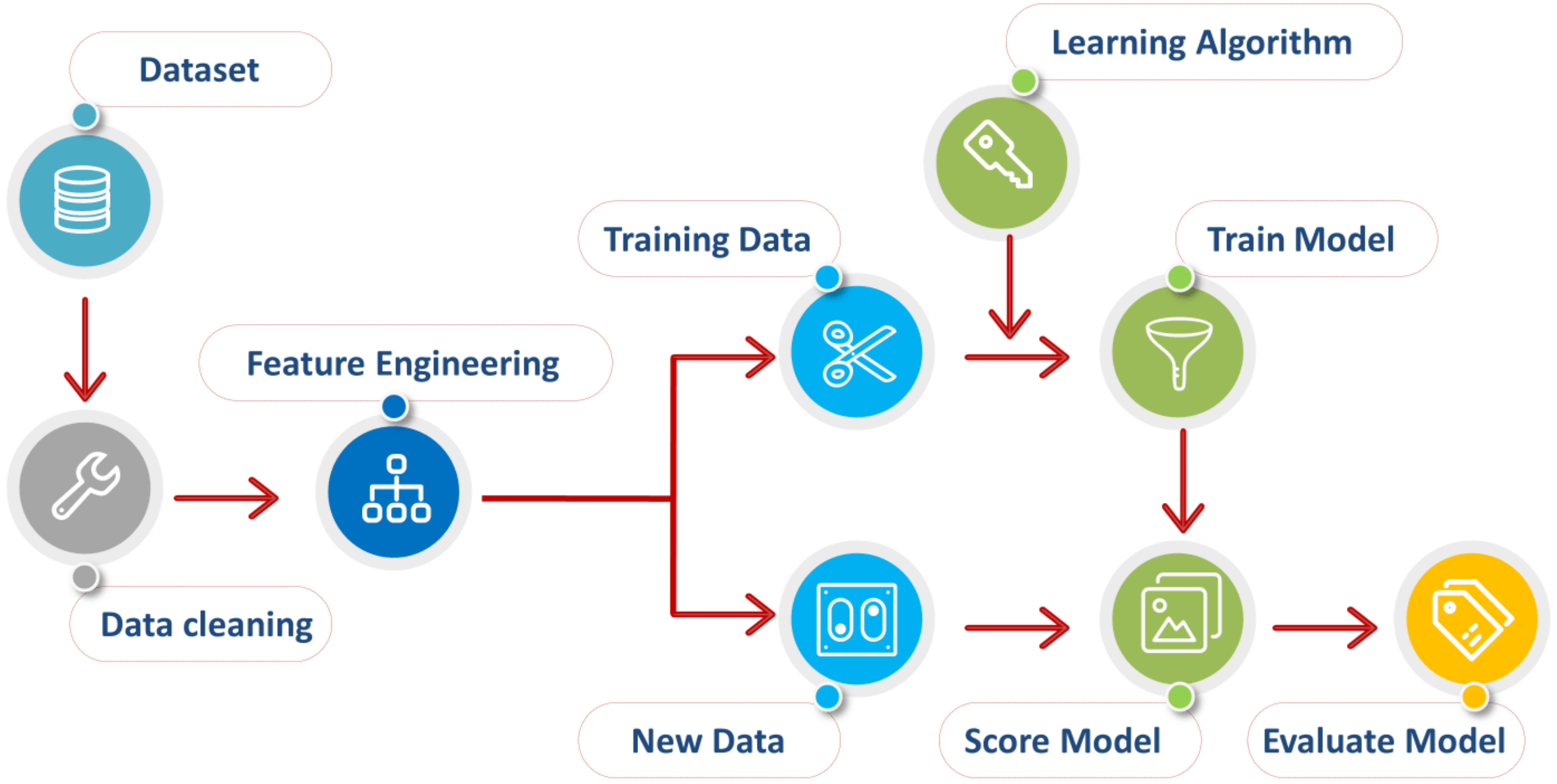
- **Research questions we would like to answer by analyzing the data**
  - Analyzing tweet metadata:
    - How do the following in tweets with misinformation compare with those in tweets with factual information:
    - Follower/Following ratio of the account.
    - Ratio of likes and account age.
    - Associated links
    - Platform

RESEARCH QUESTIONS – MACHINE LEARNING MODELS

Research questions we would like to answer by doing ML on the data

- Predict whether a tweet about COVID19 contains misinformation or not.

ML DEVELOPMENT PROCESS OVERVIEW



Source: https://medium.datadriveninvestor.com/

STAGE 1: DATA PREPROCESSING

- Deal with missing data:
  - Drop rows
  - Impute
- Preprocessing text data:
  - Cleaning to remove irrelevant items such as emojis.
  - Tokenize the text data.
  - Remove stop words.
  - Perform parts of speech tagging.
  - Perform Stemming/lemmatization.
- Preprocessing other features:
  - Reduce features by combining columns. Eg - make a new column called follower/following ratio and drop those columns.
  - One hot encode categorical data.
- Identify and remove any outlier data.

STAGE 2 – DATA LABELING

- Study how we would solve the problem manually.
- Use our own judgement to decide whether a tweet contains misinformation or not.
- For supervised learning tasks, identify the target attribute(s).

STAGE 3 – FEATURE EXTRACTION

- TF-IDF: This gives us a metric that is proportional to the frequency of occurrence of a term in a document but inversely proportional to the number of document it appears in

STAGE 4 – CLASSIFICATION

- Use a Decision tree classifier with default parameters and use that as our base model.
- Try out the following classification models with the default parameters and see how the compare to our base model:
  - Random Forest Classifier
  - SVM
  - NaiveBayes Classifier

STAGE 5 – VALIDATION OF THE ML MODEL PERFORMANCE

- Make predictions for the test dataset and use the predictions along with the true labels to get the score.
- Use the following metrics:
- F1 Score
  - Accuracy
- For each model, use N-fold cross-validation and compute the mean and standard deviation of the performance measure on the N folds.
- Shortlist the top three to five most promising models, preferring models that make different types of errors.
  - Cross-validation

STAGE 6 – OPTIMIZATION OF ML MODEL PERFORMANCE

- Choose the model that performs the best with default parameters.
- Narrow down the best possible ranges for the hyperparameters for that model
- Use Gridsearch to get the best hyperparameters.
- Ensemble - combining better performing models together.

REFERENCES

- Drishti, Jain, and Tavpritesh Sethi. n.d. "The State of Infodemic on Twitter.” Indraprastha Institute of Information Technology.