PREDICTING POLITICAL PARTISANSHIP ON SOCIAL MEDIA

Rafaëlle Aygalenq - Sarah Lina Hammoutene - Dora Linda Kocsis - Noémie Quéré

Overview

59% of Americans prefer to read news online

Individual influencers use these channels to weigh in on the way we form our opinions.

The network surrounding tweets with a political message substance shows a highly partisan structure with low connectivity among users from opposing sides

Social media play a key role in transforming political dialogue across the United States.

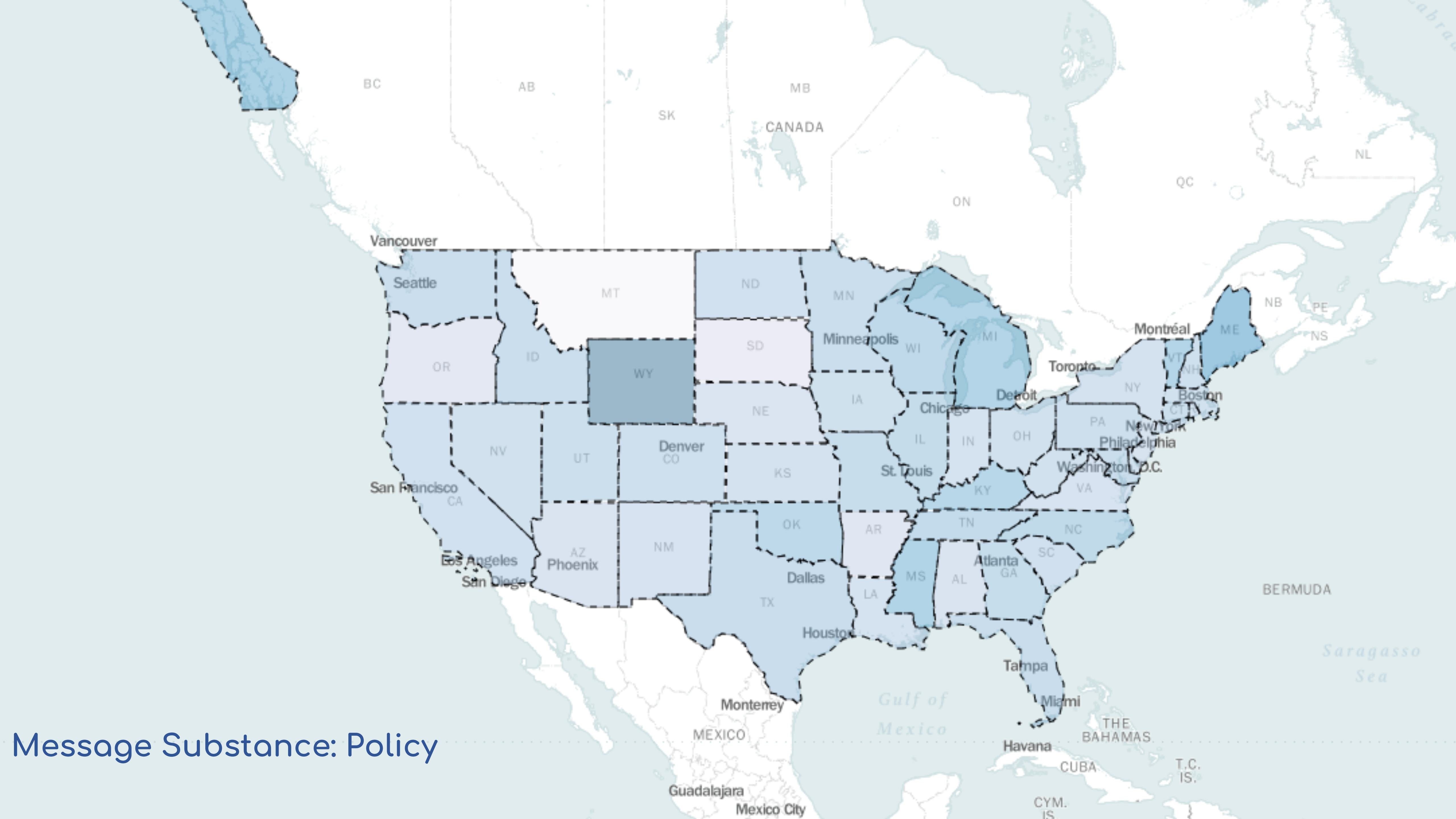
Social networks

Twitter

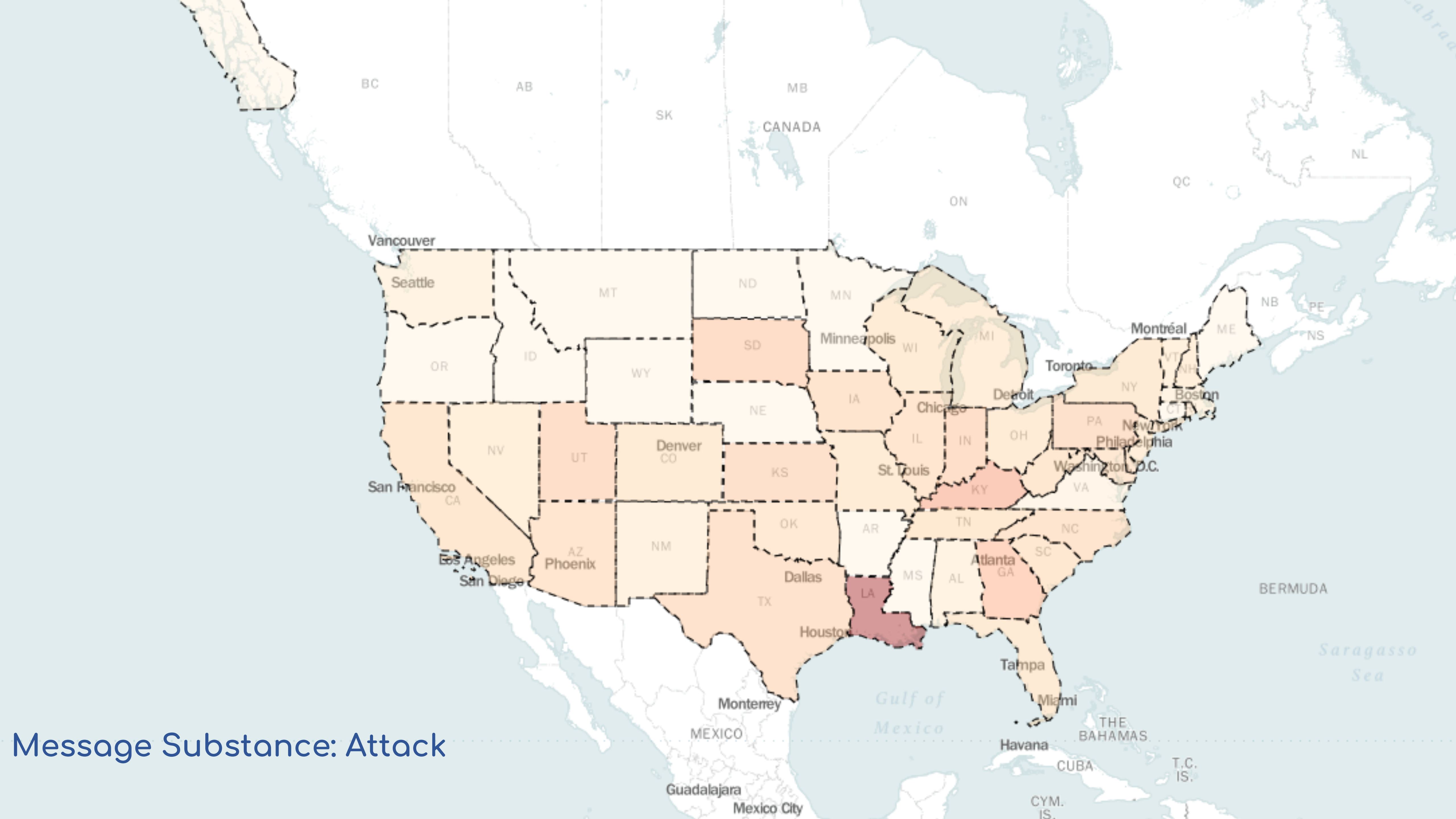
Facebook

Goal

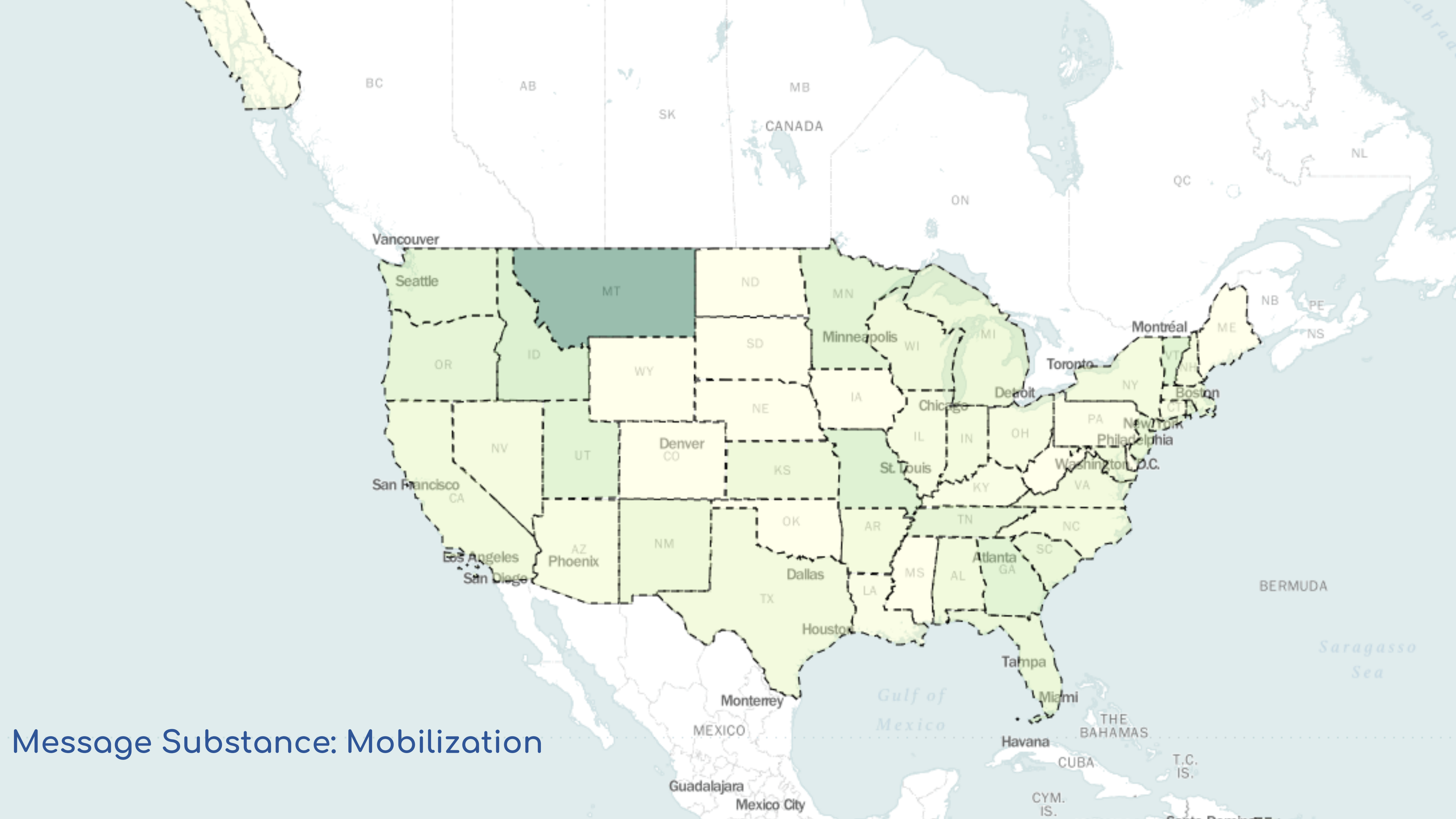
Build a model to examine whether different message substances of social media posts from US legislators convey partisan bias or not



Message Substance: Policy



Message Substance: Attack



Message Substance: Mobilization

Our Dataset

From CrowdFlower & govtrack.us

5K tweets and Facebook messages

29 features

9 types of message substance

Our Dataset

From CrowdFlower & govtrack.us

5K tweets and Facebook messages

From US Senators and other American politicians

29 features

9 types of message substance

Our Dataset

From CrowdFlower & govtrack.us

5K tweets and Facebook messages

29 features

- human judgments on the intended audience
- existence of partisan bias
- author information (state, party, gender, etc).
- types of message substance.

9 types of message substance

Our Dataset

From CrowdFlower & govtrack.us

5K tweets and Facebook messages

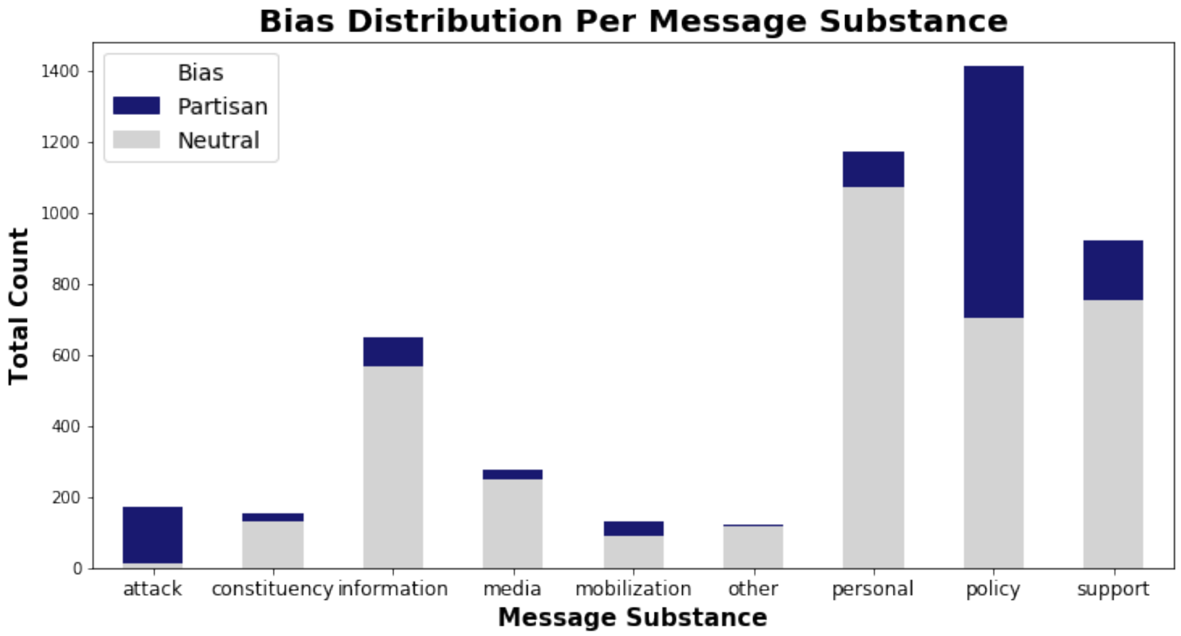
29 features

9 types of message substance

informational, support, media appearance, attack, policy, personal, mobilization, other

PRELIMINARY ANALYSIS

WordCloud



Bias/ Message Substance

PIPELINE

DATA COLLECTING

Merge two different datasets

Feature engineering

DATA PRE-PROCESSING

Cleansing

Text processing

Bag of words model

SENTIMENT ANALYSIS

Message Substance

Bias

PREDICTION

EVALUATION

- Accuracy
- Precision
- Recall
- AUC Score

TEXT PRE-PROCESSING

TOKENIZATION

Process of breaking up sentences into smaller pieces

STOP WORDS REMOVAL

Filtering out commonly used words before indexing the entries by their frequencies

STEMMING

Process of obtaining the base form of a word by reducing any affixes

BAG OF WORDS MODEL

Process of segmenting sentences into words building a unique dictionary for the dataset in place

MODELS OVERVIEW

Logistic Regression

Belongs to the Generalized Linear Models family

K Nearest Neighbors

Computation of a similarity measure across observations

Naïve Bayes Classifier

Probabilistic classifier based on Bayes theorem

Support Vector Machines

Research of the optimal margin hyperplanes

Random Forest

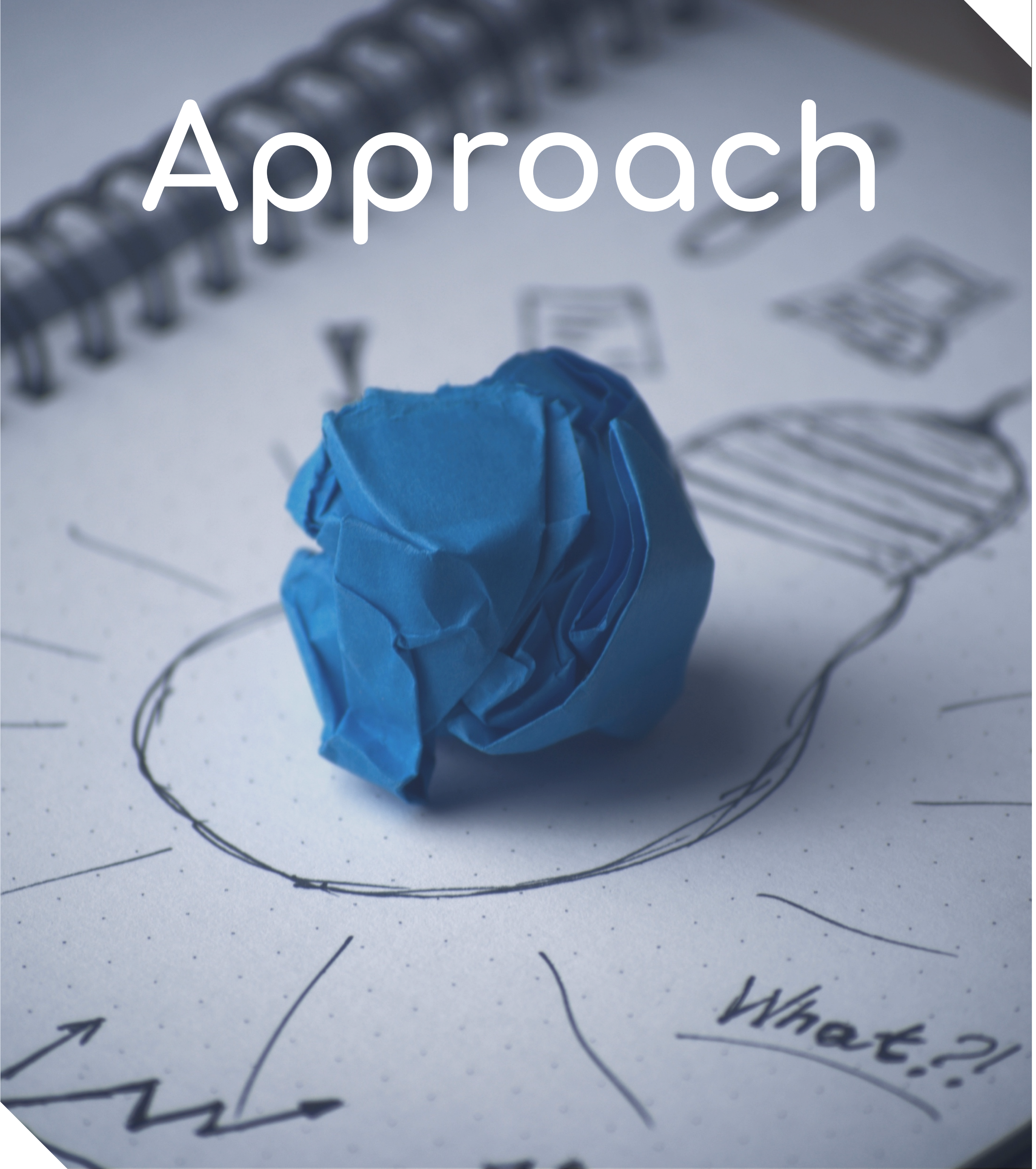
Ensemble machine learning method which develops multiple decision trees

Neural Network

Association into a graph more or less complex of artificial neurons

Boosting

Construction of a family of models aggregated together in an adaptive way : weak learners

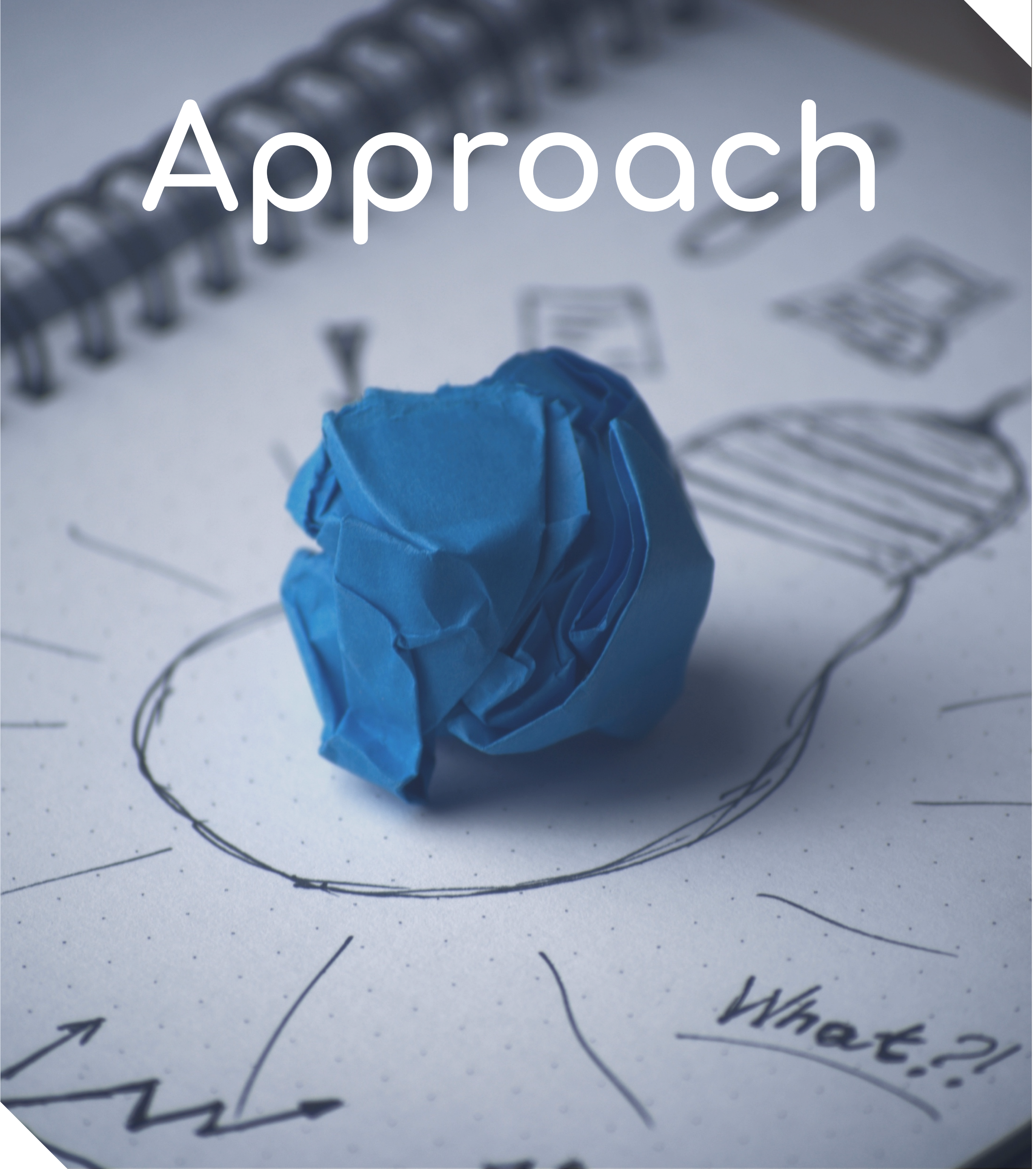


MESSAGE SUBSTANCE

Multiclass Classification Problem

9 categories in one model

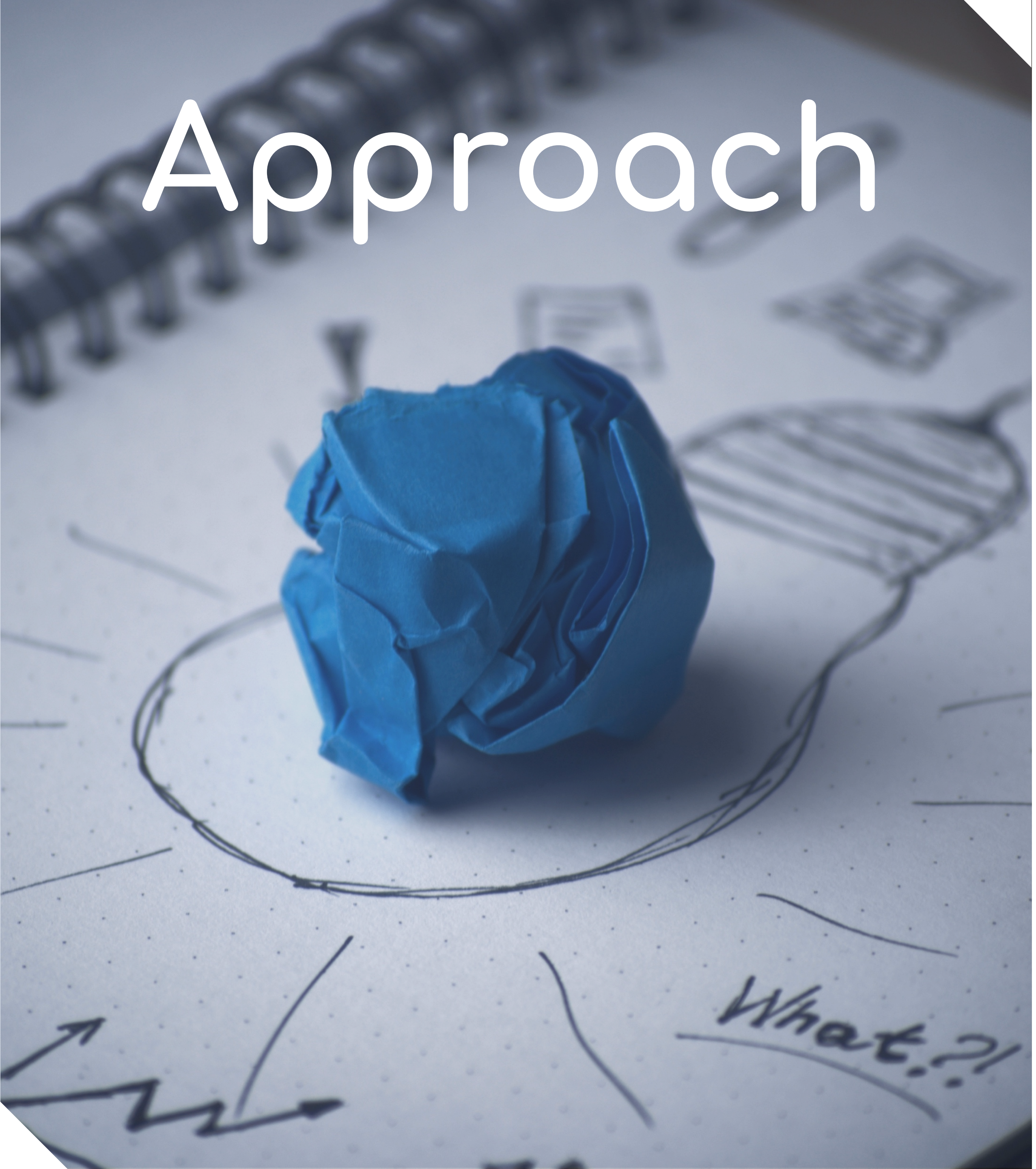
Approach



MESSAGE SUBSTANCE

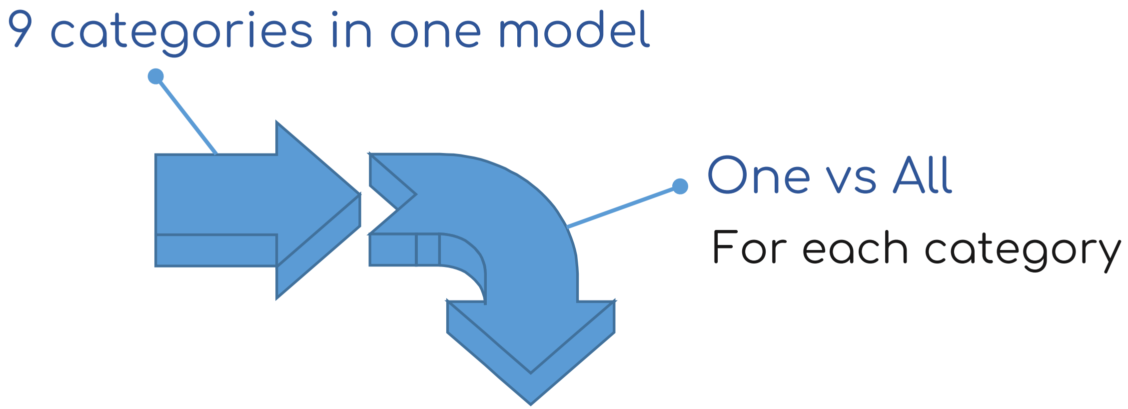
Multiclass Classification Problem

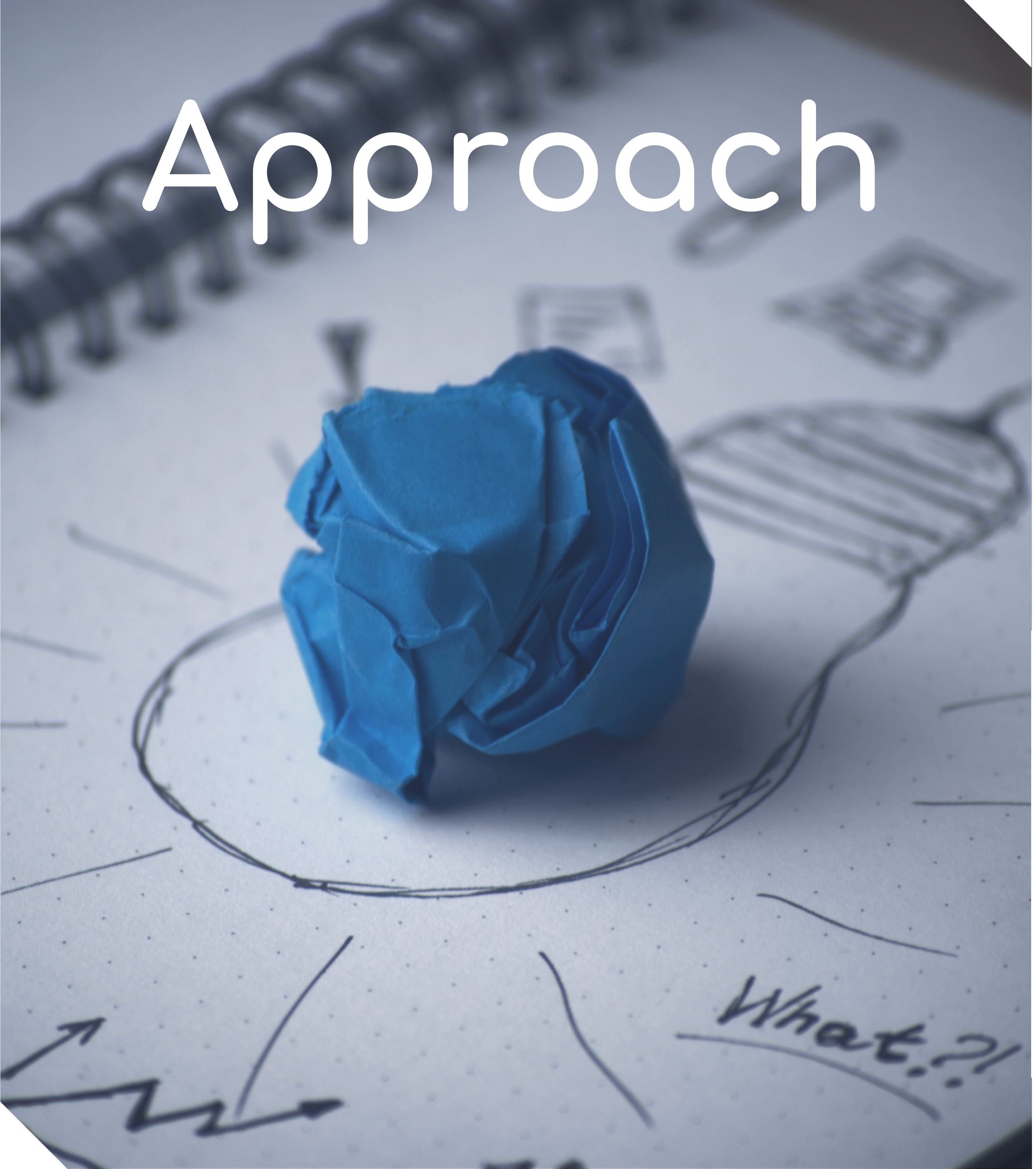
```
F1       [ 0.3043022   0.          0.06926407  0.          0.          0.17486339
  0.          0.04878049  0.        ]
Precision[ 0.20596591  0.          0.20512821  0.          0.          0.23529412
  0.          0.04210526  0.        ]
Recall   [ 0.58232932  0.          0.04166667  0.          0.          0.13913043
  0.          0.05797101  0.        ]
Accuracy 0.189
```



MESSAGE SUBSTANCE

Multiclass Classification Problem



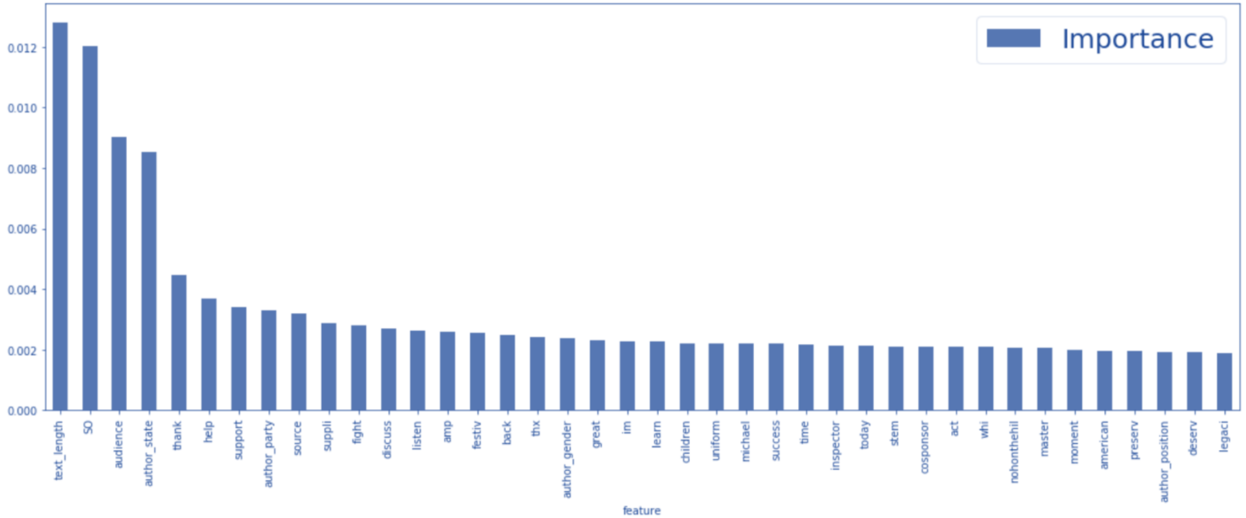


MESSAGE SUBSTANCE

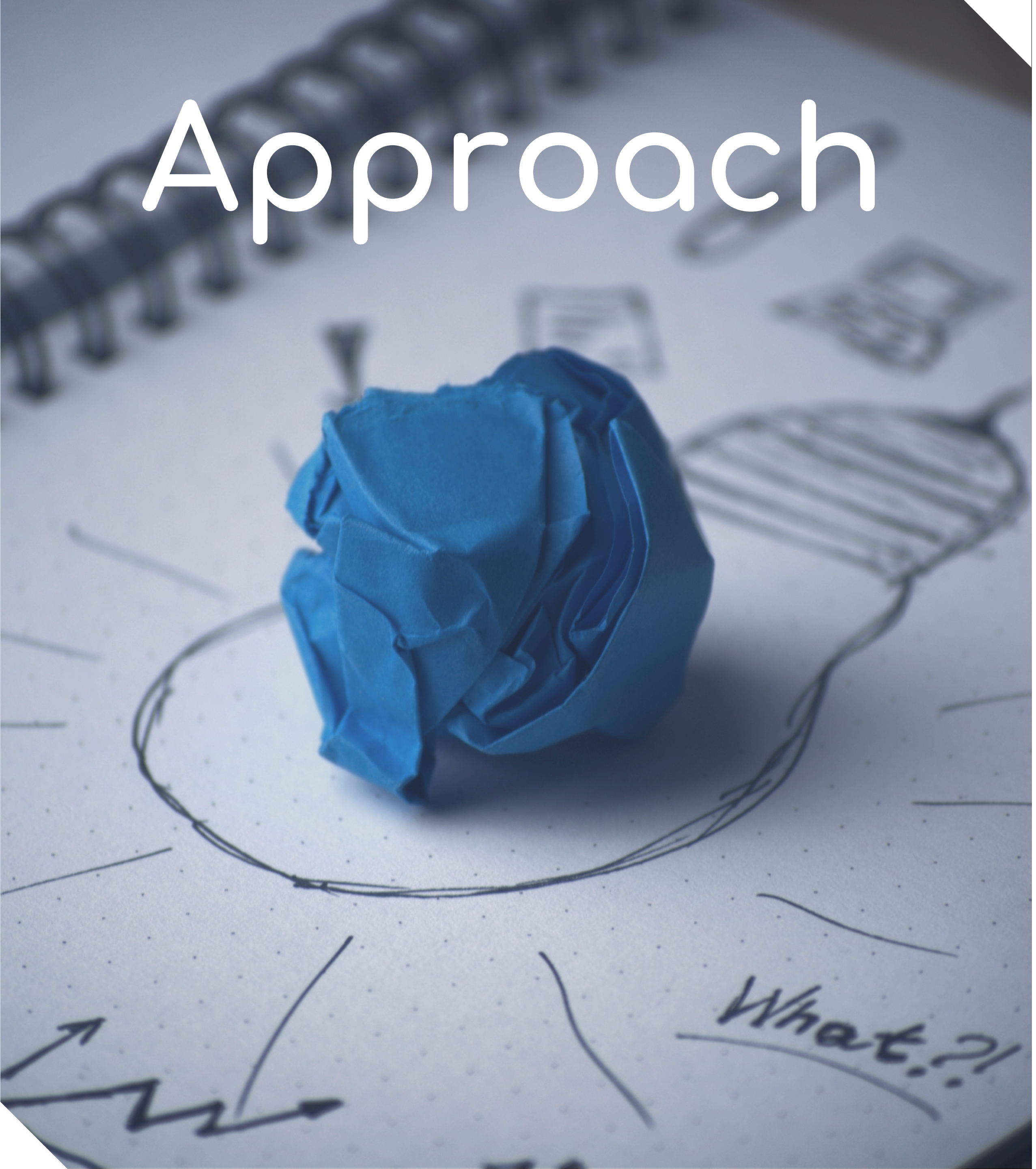
Multiclass Classification Problem

```
====================================
Testing RandomForestClassifier
Learing time 1.66872310638s
Predicting time 0.0592069625854s
====================================
F1         [ 0.88851728  0.02912621]
Precision[ 0.80831643  0.21428571]
Recall     [ 0.98638614  0.015625   ]
Accuracy 0.8
====================================
```

Results For The Category "Support"



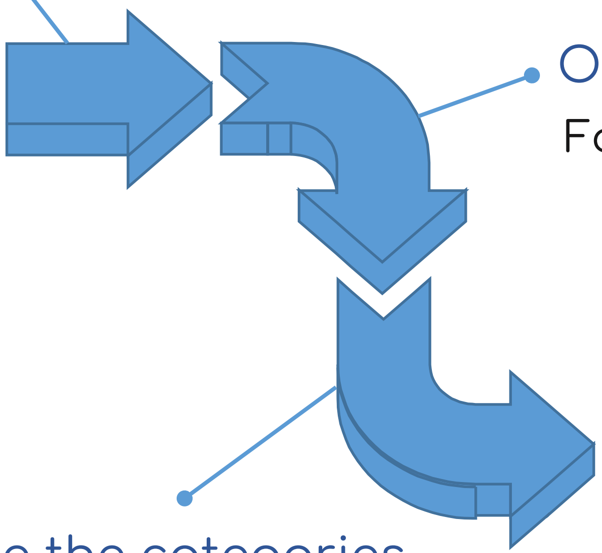
Approach



MESSAGE SUBSTANCE

Multiclass Classification Problem

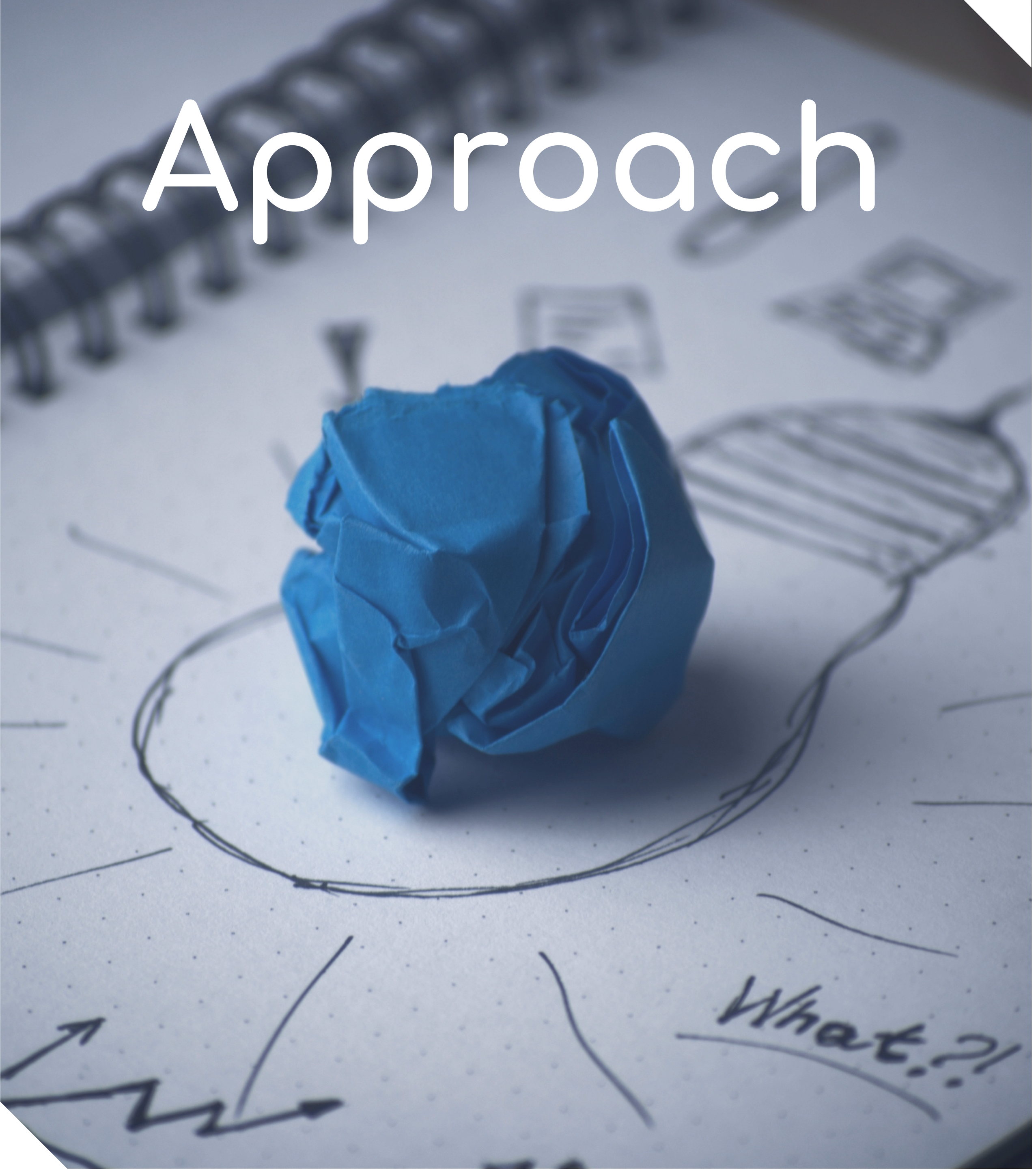
9 categories in one model



One vs All

For each category

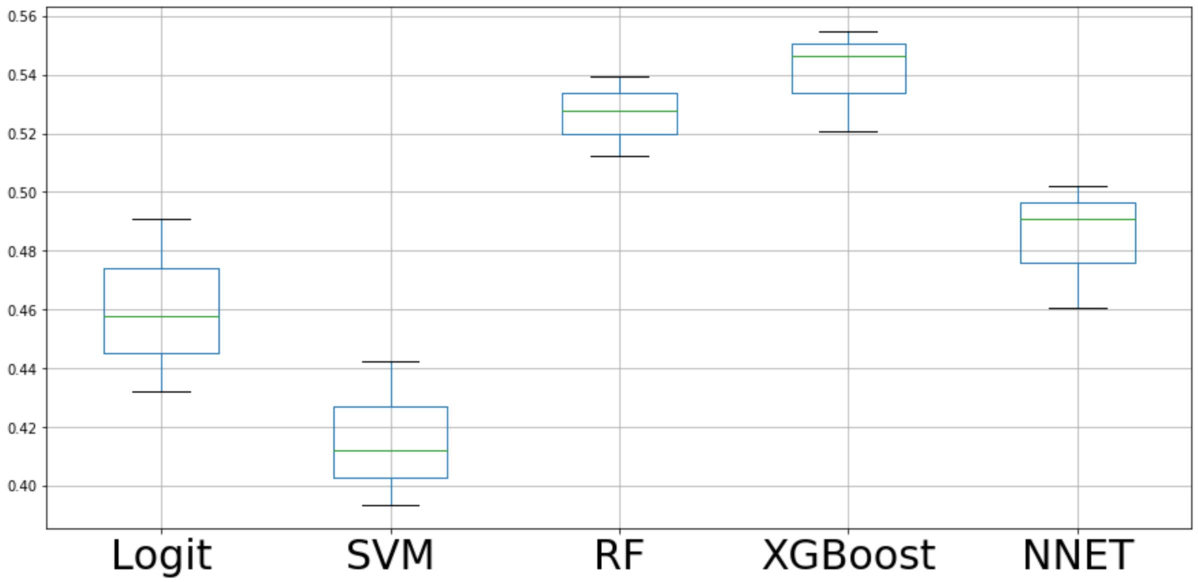
Merge the categories
into 3 or 4 groups

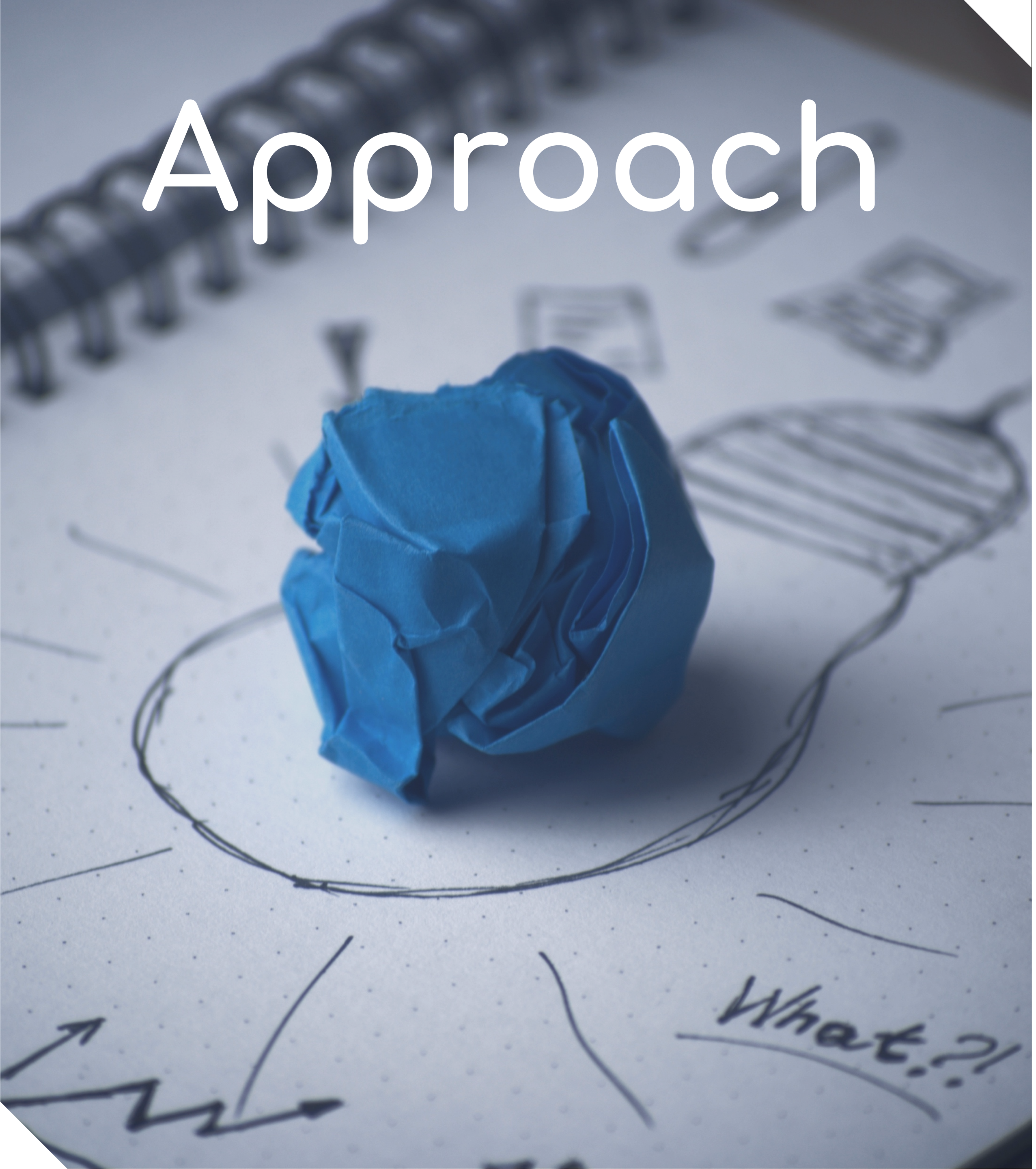


MESSAGE SUBSTANCE

Multiclass Classification Problem

Category 0 : keep the actual category 0
Category 1 : personal + other
Category 2 : attack + support + mobilization + constituency

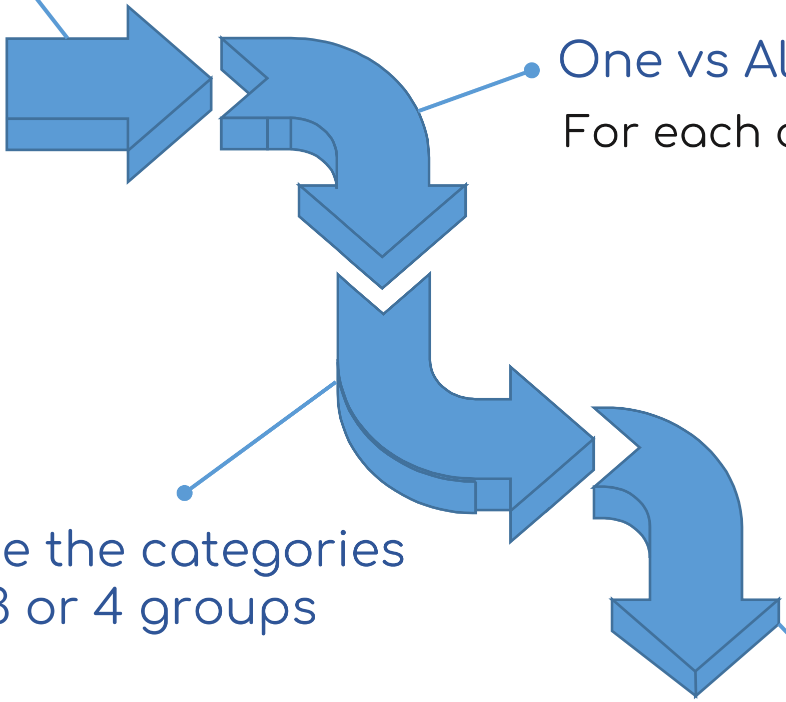




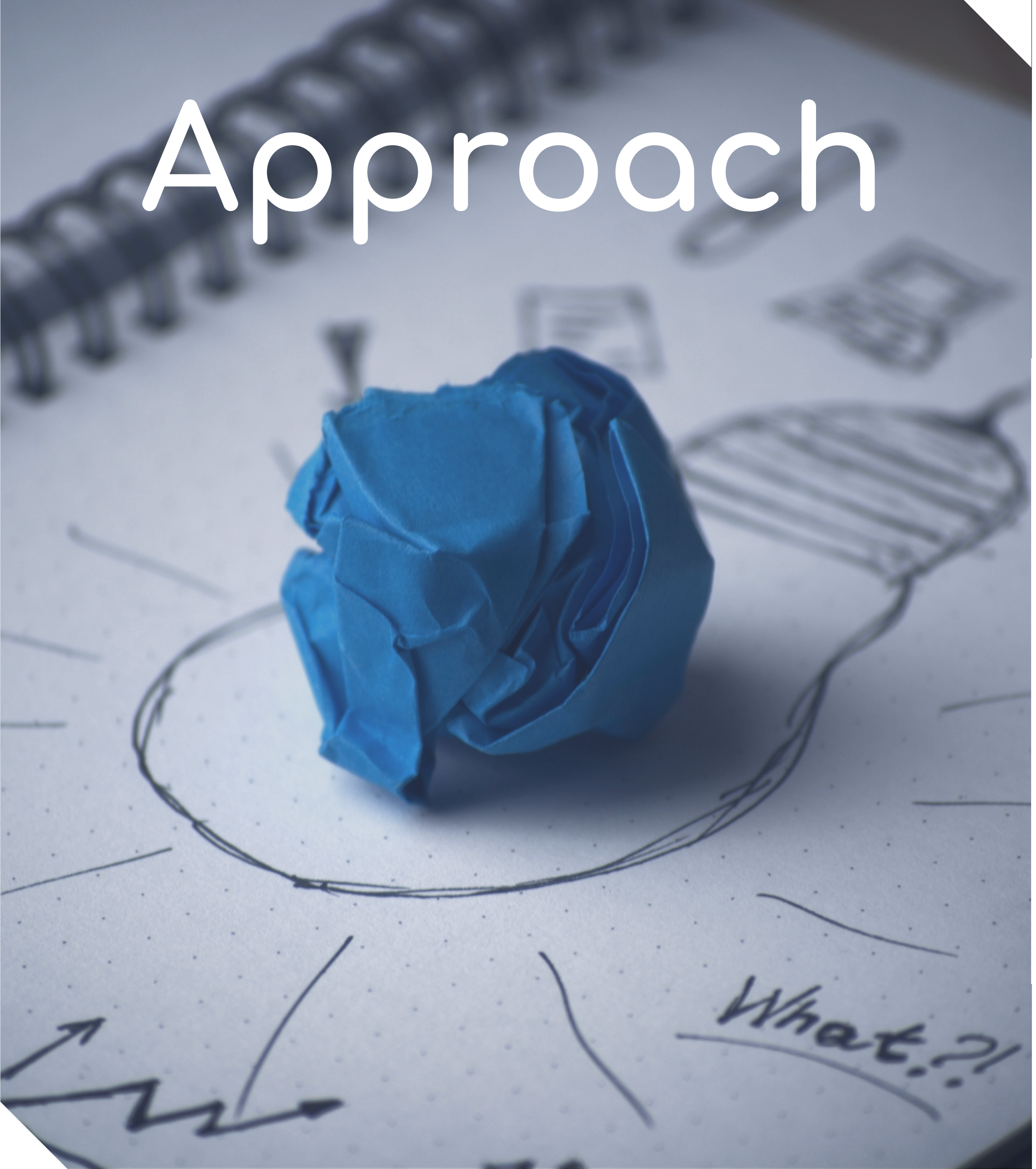
Approach

MESSAGE SUBSTANCE

Multiclass Classification Problem

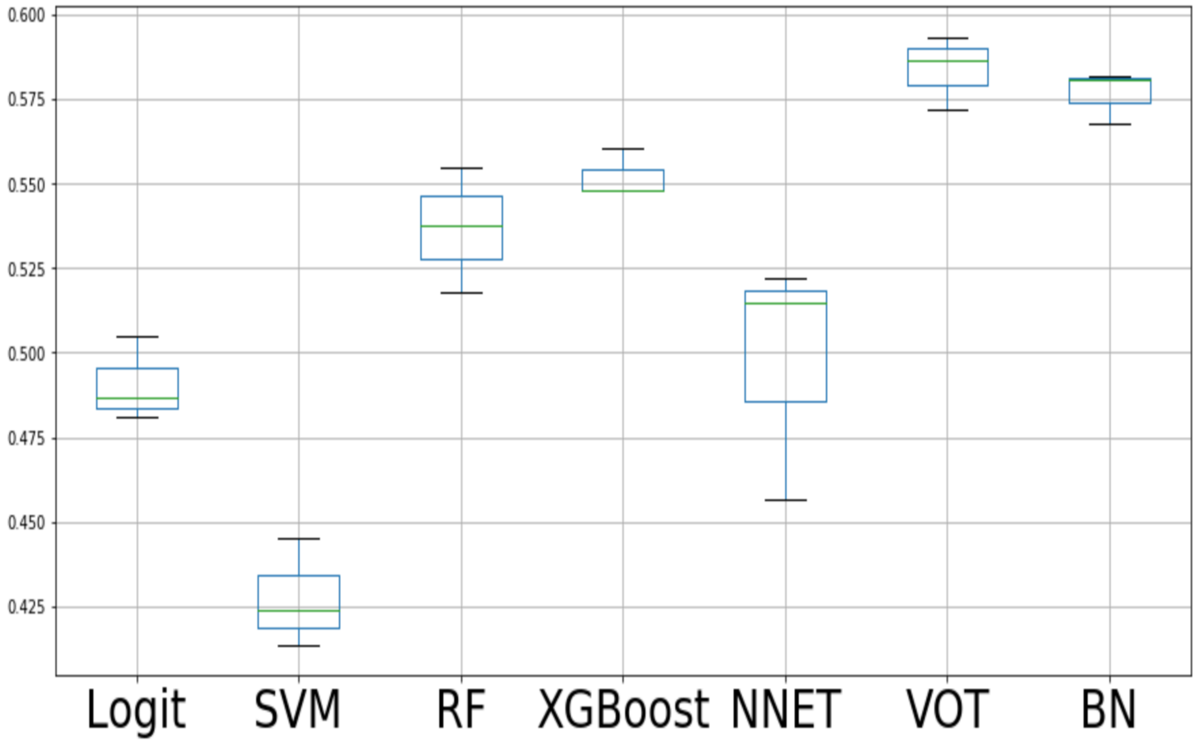


Approach



MESSAGE SUBSTANCE

Multiclass Classification Problem



BIAS

BINARY CLASSIFICATION

Determining if a social media post is biased or not

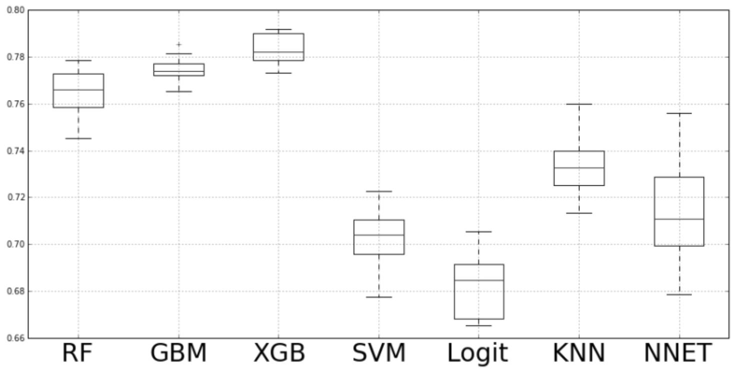
MODELING

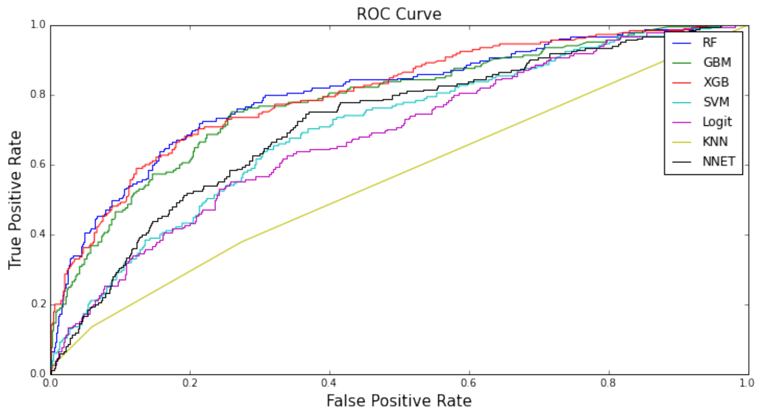
Applying different models and use cross-validation to optimize parameters

AGGREGATION

Use of the Voting Classifier which combines 3 best models (RF, GBM, XGBoost)

RESULTS





1

LIMITS OF DATASET

5000 observations for 9 classes is not enough.
More data would have help to better train the different models

2

LIMITS OF COMPUTING POWER

Could have obtain better results with other techniques that required more powerful computers
Convolutional networks used through Keras library were promising