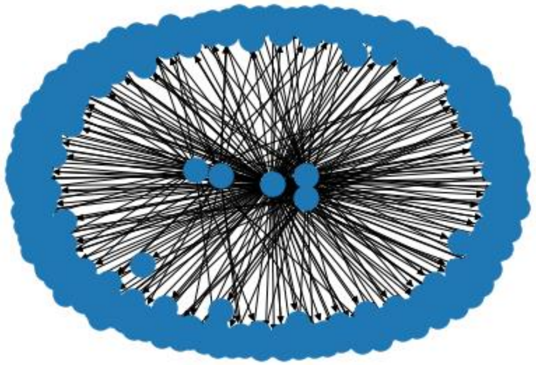
An Investigation of Fake News Classification via Hybrid Feature Selection

OLUMIDE OMOLOLU

The Case Study



This project aims to evaluate the impact of propagation data on fake news classification using baseline natural language processing methods within a social media context in a manner that is robust as well as computationally inexpensive.

Objectives of Similar Studies

Data

- Content
- Writing Style
- Interactions
- Propagation Path

Methods

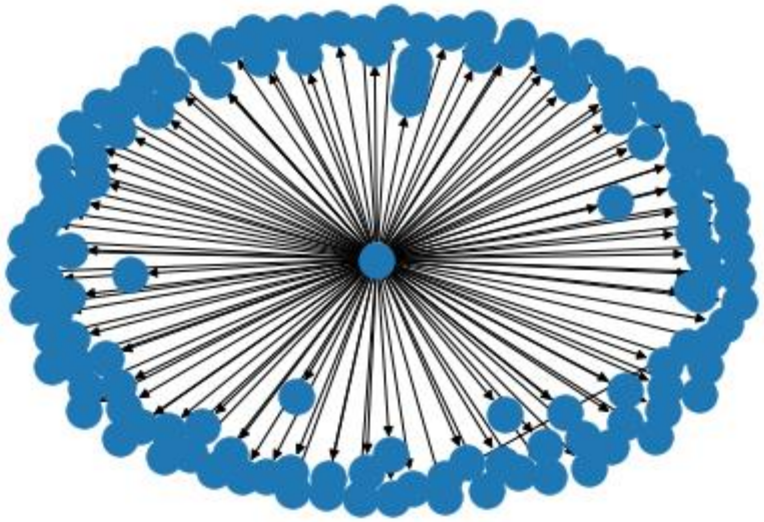
- Supervised learning
  - Deep learning

Research Aims

- Early Detection
- Malicious Account Detection

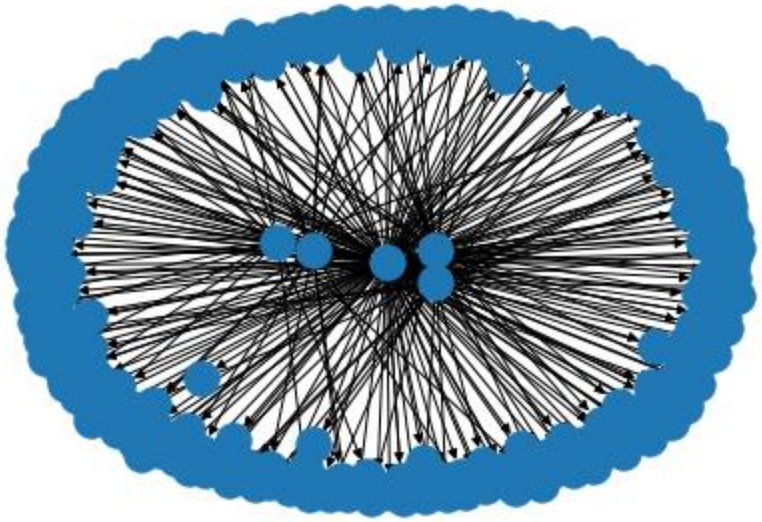
Relevance of Propagation

Information dispersal on social media can be represented as a graph or directed acyclic tree which is unique to each item of news. These graphs can be represented using a handful of metrics that define the relation of the root node with the other nodes.

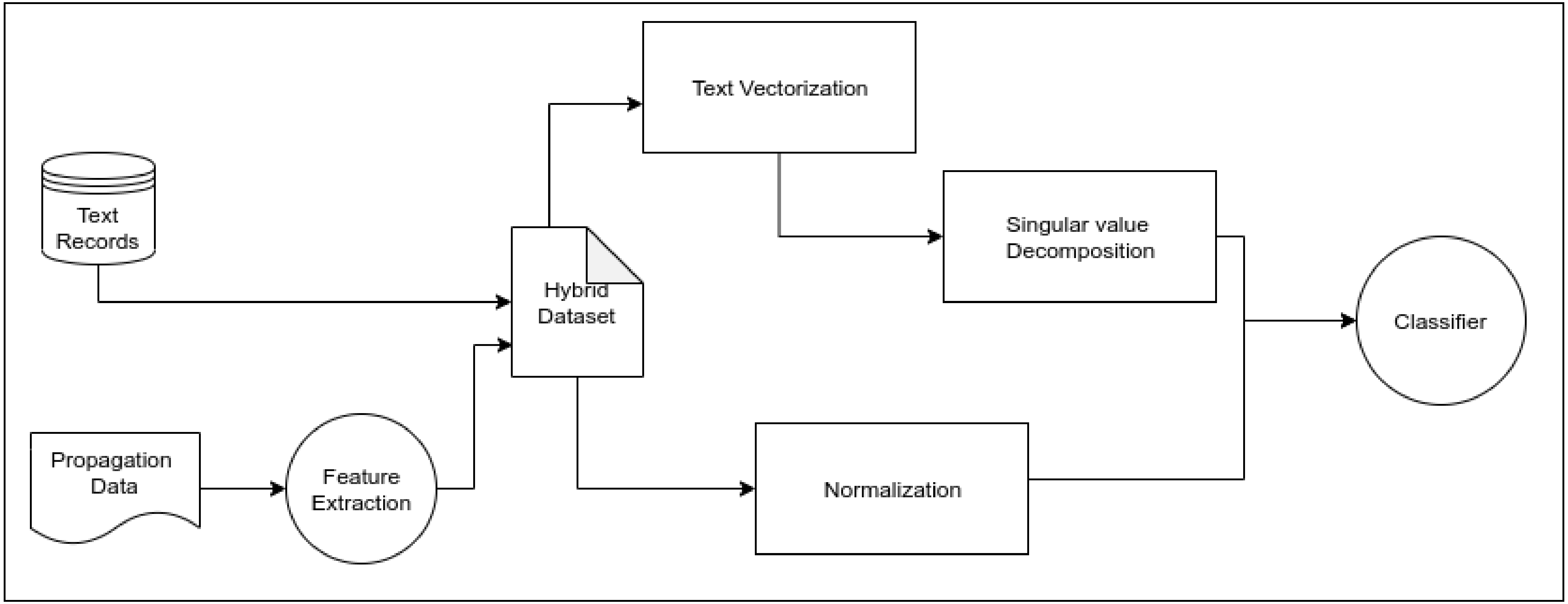


Representive Features:

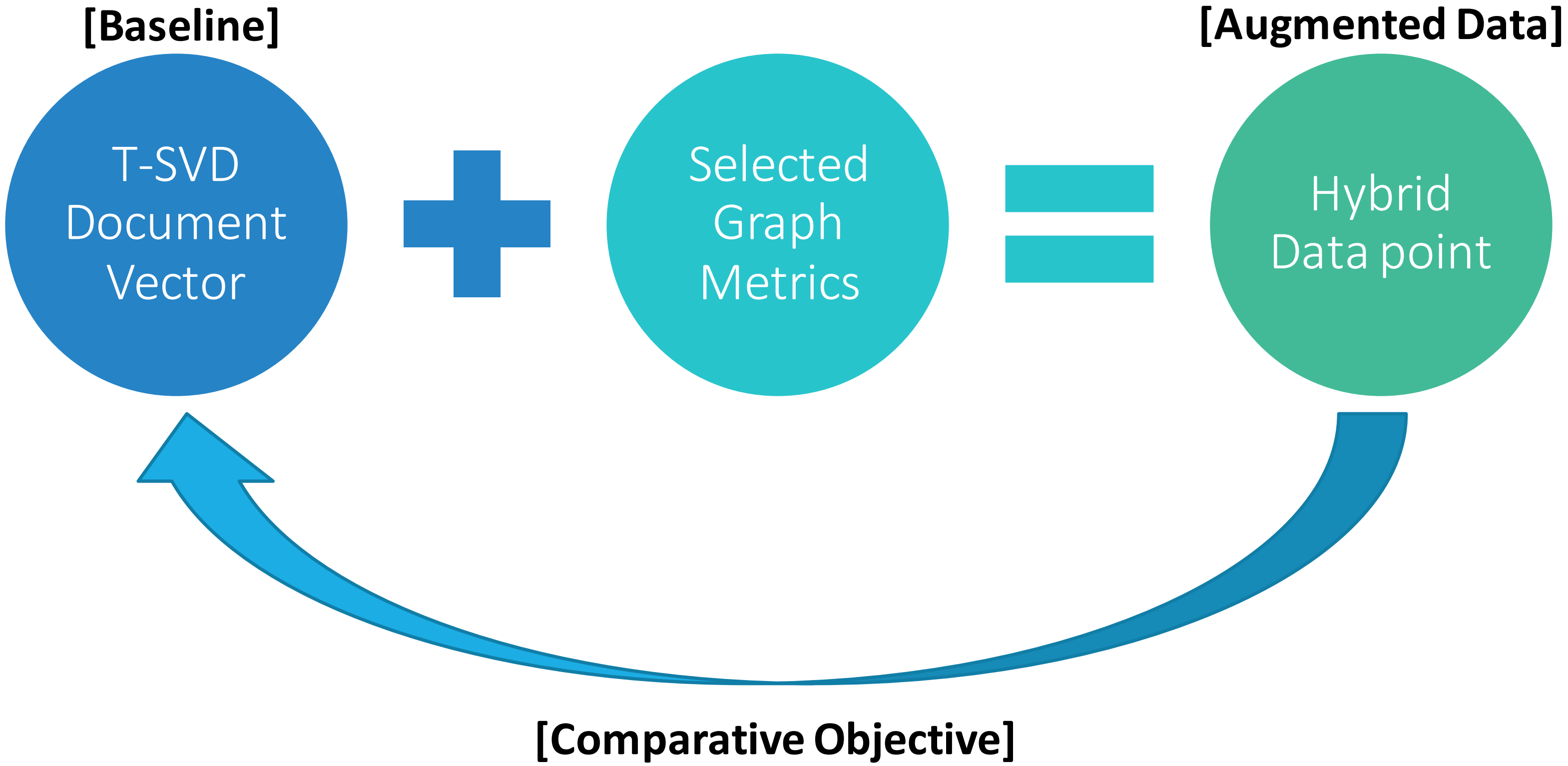
- Density
- Degree
- Centrality



Approach: Pipeline Architecture



Approach: Experimental Data Structure



Data Transformation

Document Vectorization

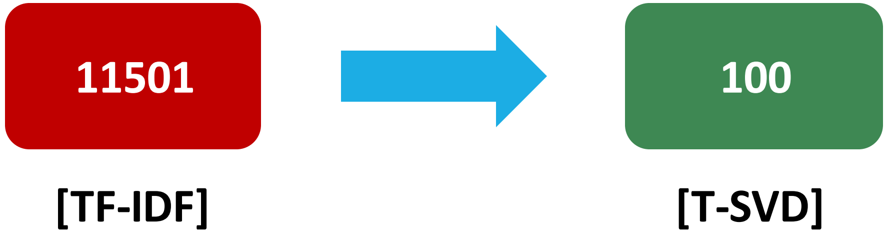
- Bag of Words Representation
- TF-IDF (Term Weighting)

Normalization

- Improvement to classifier performance
- Applicability across classifiers

Truancated Singular Value Decomposition

- Dimensionality
- Sparsity



11501

[TF-IDF]

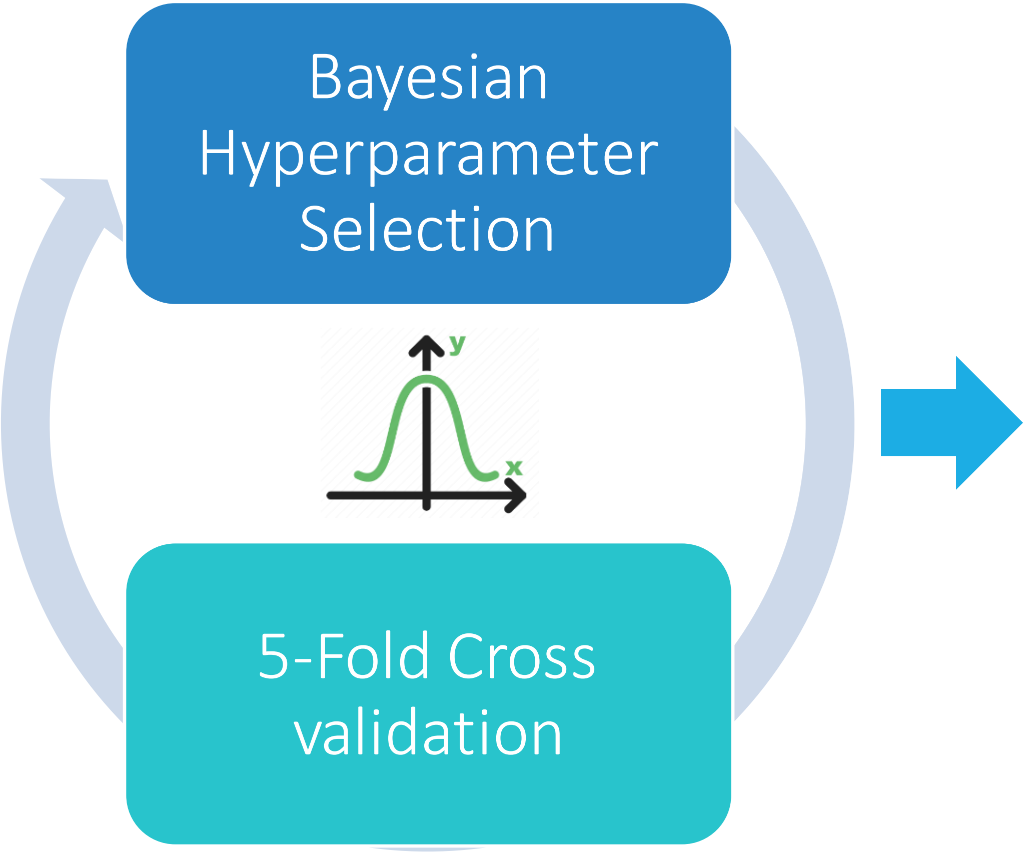
100

[T-SVD]

Bayesian Hyperparameter Tuning

Classification Methods

- Logistic Regression
- Support Vector Machines
- Naïve Bayes
- KNN
- Random Forest



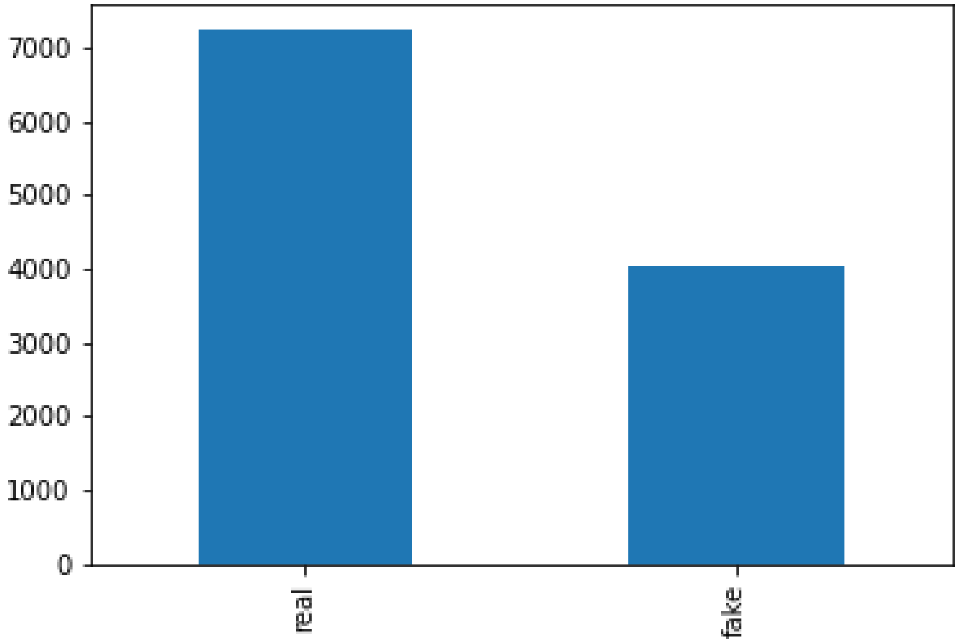
Bayesian Hyperparameter Selection

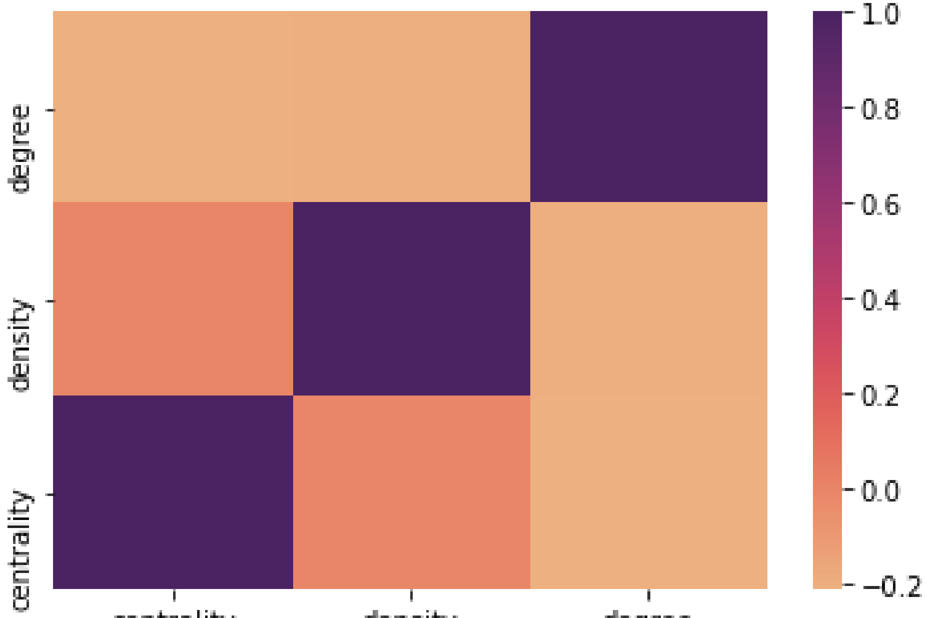
5-Fold Cross validation

[Best Performing Model]

The Data

The **FakeNewsNet** dataset consists of two datasets which are both madeup of linguistic and temporal dispersion data about posts on Twitter, from well known fact checking websites. 11,259 entries from the combined dataset were utilized in this study.

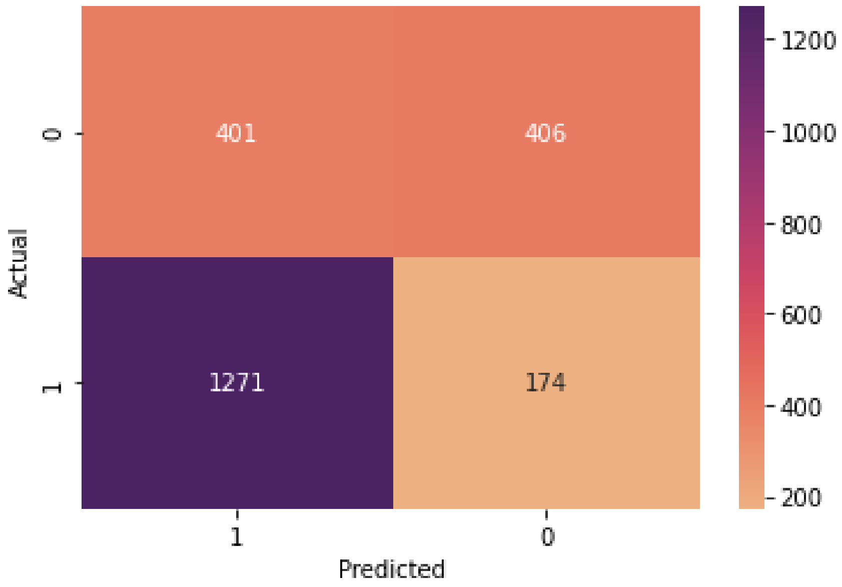




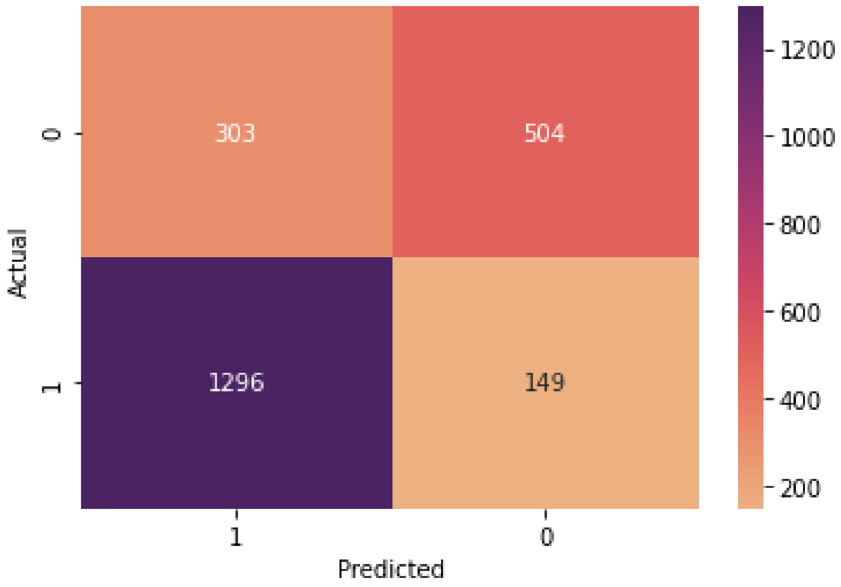
Comparative Review

Logistic Regression Example

Text-Only



Combined Approach



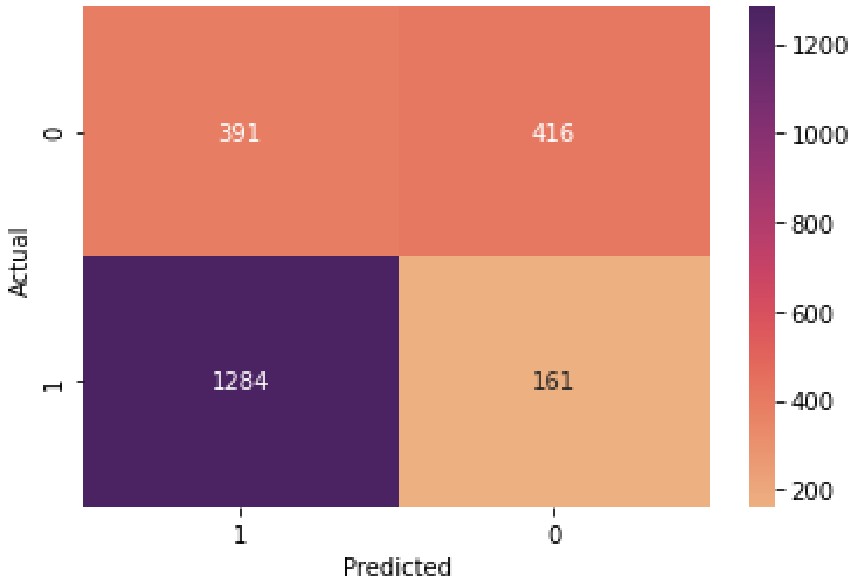
Comparative Review & Statistical Testing

Classifier	Accuracy		F1-Score		P-Value
	Hybrid	Text	Hybrid	Text	
Logistic Regression	0.7993	0.7447	0.8515	0.8155	9.59E-06
Naive Bayes	0.7851	0.7615	0.8373	0.834	3.03E-62
KNN	0.7882	0.7376	0.8458	0.8163	2.43E-11
SVM	0.8193	0.7411	0.8656	0.8163	7.95E-15
Random Forest	0.8393	0.7549	0.8782	0.8231	6.37E-14

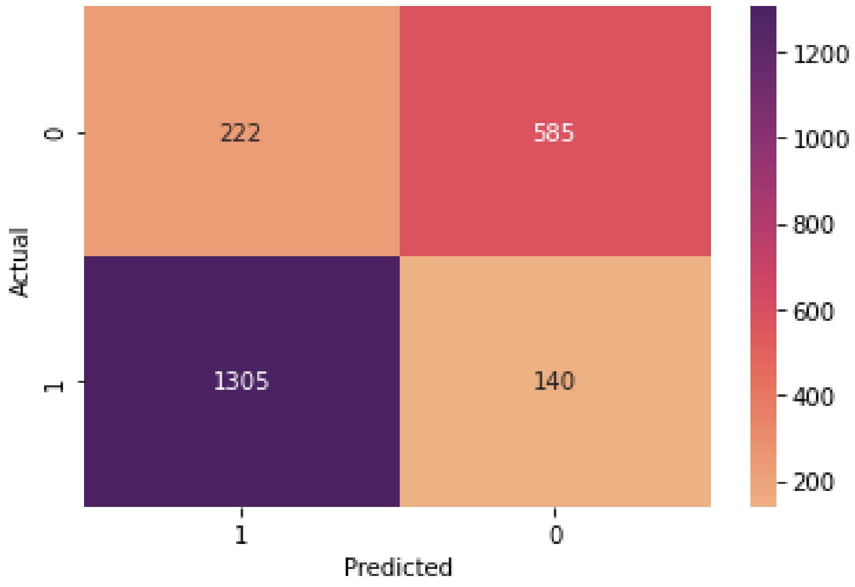
Comparative Review: Detector vs Filter

Random Forests Example

Text-Only



Combined Approach



Reflections

Where to go Next

- Early Detection
- Feature Investigation
- Incorporation of other signals
  - Time
  - User interaction
- Implication for deep learning approaches

Challenges/ Areas for Improvement

- Data Availability
- Computational Resources
- Depth of Text and Graphical processing