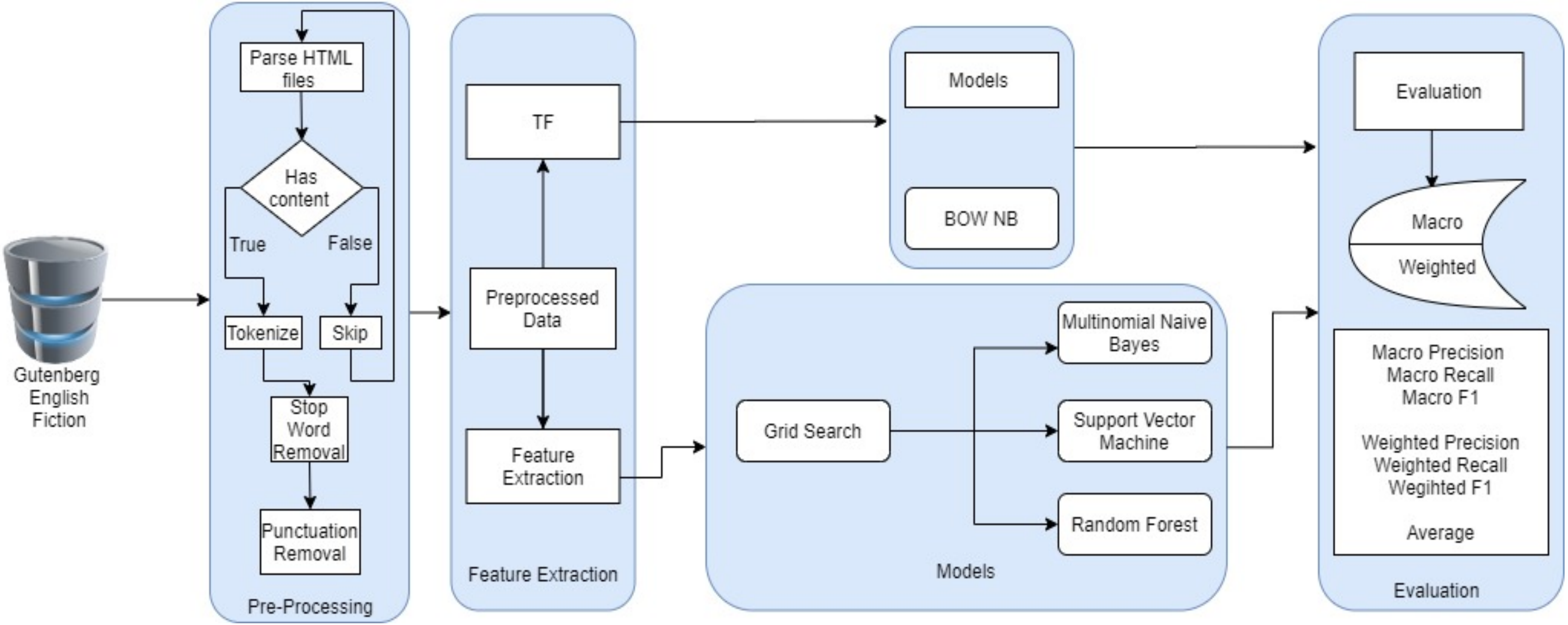
Genre Identification on a subset of Gutenberg Corpus

Project Team Members:

Himanshi Bajaj - 225827
Nandish Bandi Subbarayappa - 229591
Steve Simon - 229497
Sujith Nyarakkad Sudhakaran - 229879

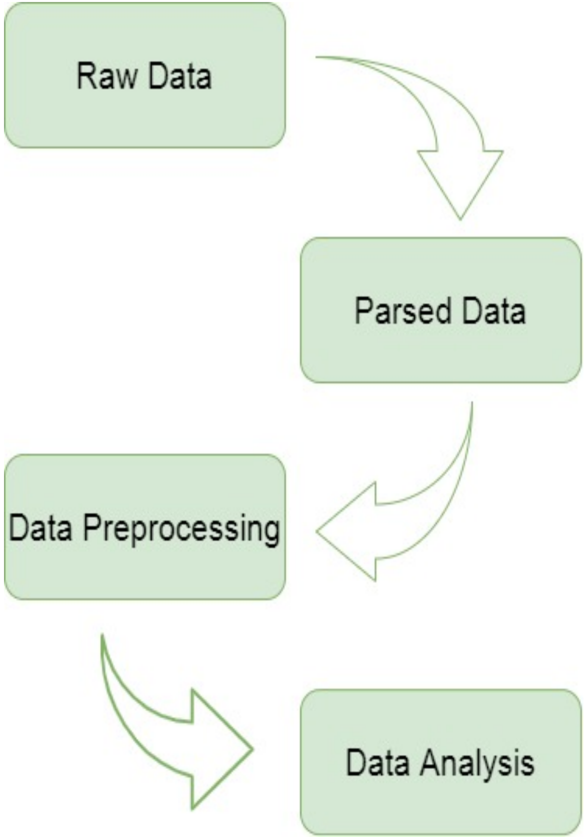
Advanced Topics in Machine Learning - Summer 2020

Overview


Gutenberg English Fiction
Parse HTML files
Has content
True
False
Tokenize
Skip
Stop Word Removal
Punctuation Removal
Pre-Processing
TF
Preprocessed Data
Feature Extraction
Feature Extraction
Models
BOW NB
Grid Search
Multinomial Naive Bayes
Support Vector Machine
Random Forest
Models
Evaluation
Macro
Weighted
Macro Precision
Macro Recall
Macro F1
Weighted Precision
Weighted Recall
Wegihted F1
Average
Evaluation

Preprocessing Steps:

- Parse HTML files to remove <p> tags
- Check for documents with no content
- Do tokenization for all features, but perform punctuation or stop word removal for certain features
- Reduced the text content to save the processing time



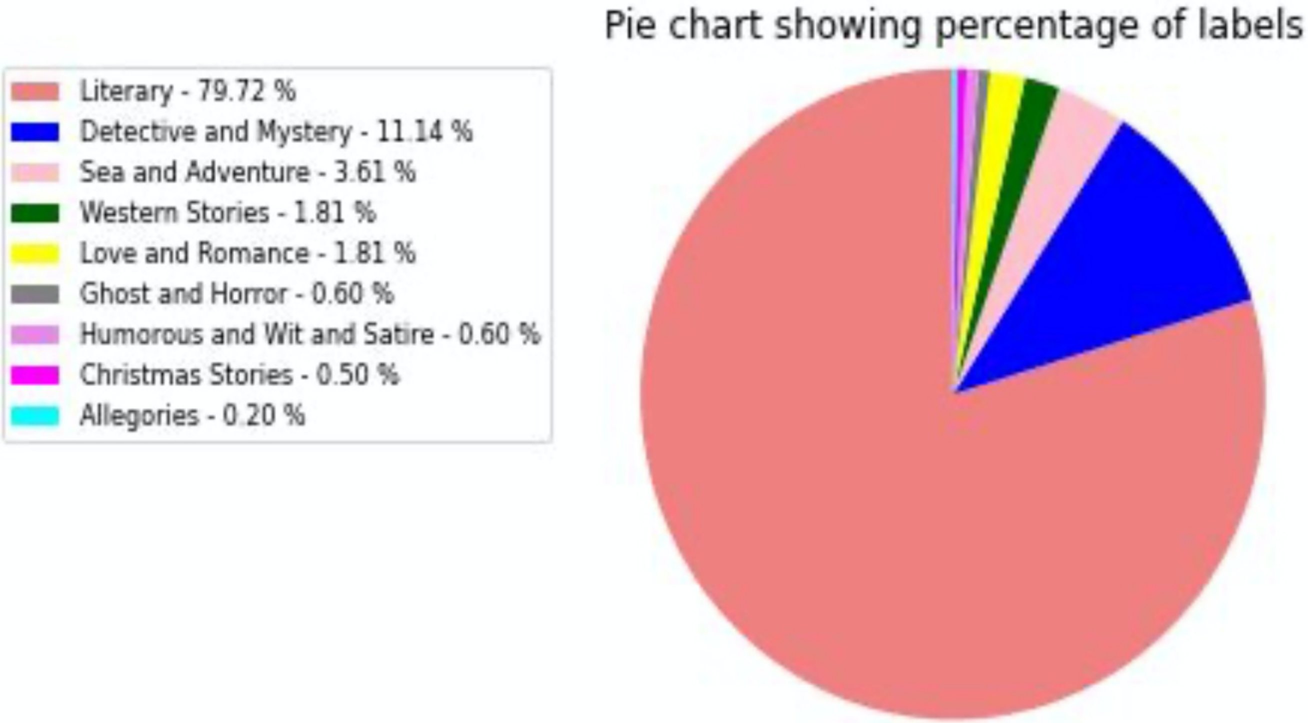
Raw Data

Parsed Data

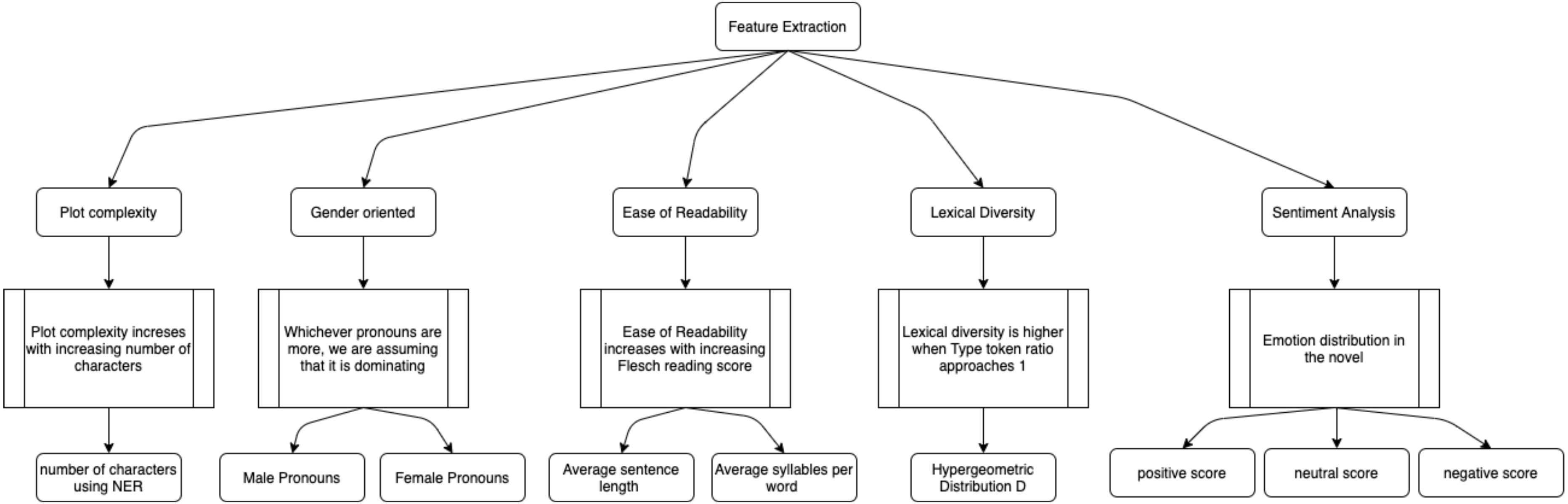
Data Preprocessing

Data Analysis

Data Analysis



Features Extracted


Feature Extraction

Plot complexity

Plot complexity increses with increasing number of characters

number of characters using NER

Gender oriented

Whichever pronouns are more, we are assuming that it is dominating

Male Pronouns

Female Pronouns

Ease of Readability

Ease of Readability increases with increasing Flesch reading score

Average sentence length

Average syllables per word

Lexical Diversity

Lexical diversity is higher when Type token ratio approaches 1

Hypergeometric Distribution D

Sentiment Analysis

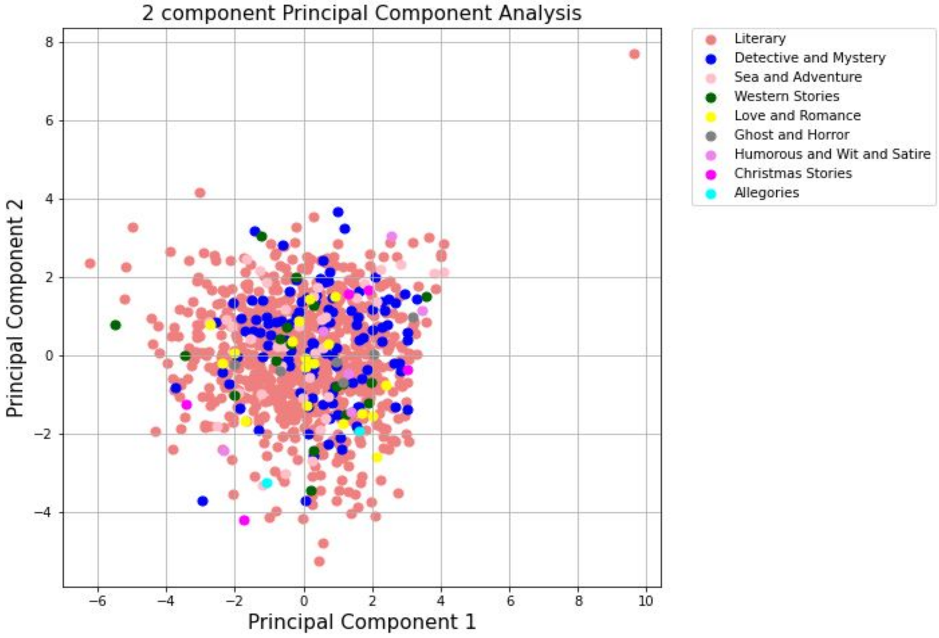
Emotion distribution in the novel

positive score

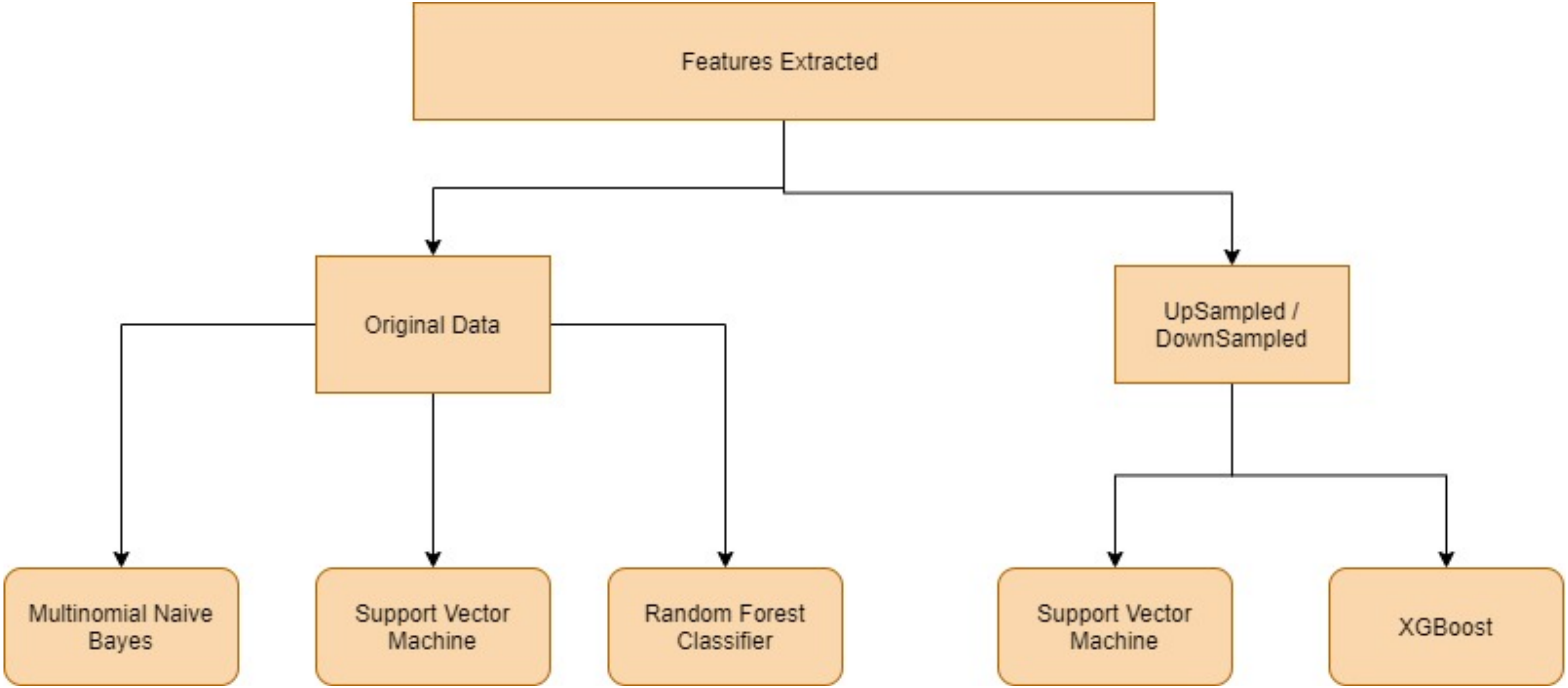
neutral score

negative score

Features Analysis



Models



Features Extracted

Original Data

UpSampled / DownSampled

Multinomial Naive Bayes

Support Vector Machine

Random Forest Classifier

Support Vector Machine

XGBoost

Evaluations - 1

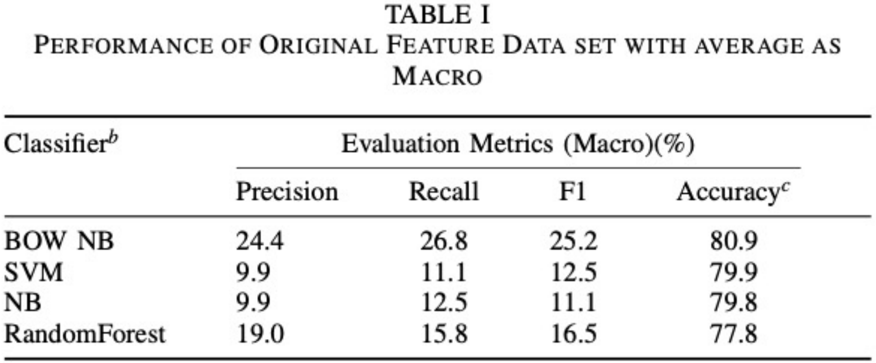
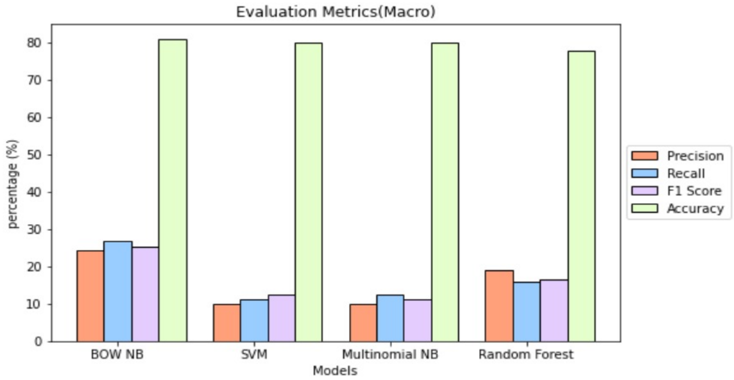


TABLE I
PERFORMANCE OF ORIGINAL FEATURE DATA SET WITH AVERAGE AS MACRO

Classifier[b]	Evaluation Metrics (Macro)(%)			
	Precision	Recall	F1	Accuracy[c]
BOW NB	24.4	26.8	25.2	80.9
SVM	9.9	11.1	12.5	79.9
NB	9.9	12.5	11.1	79.8
RandomForest	19.0	15.8	16.5	77.8



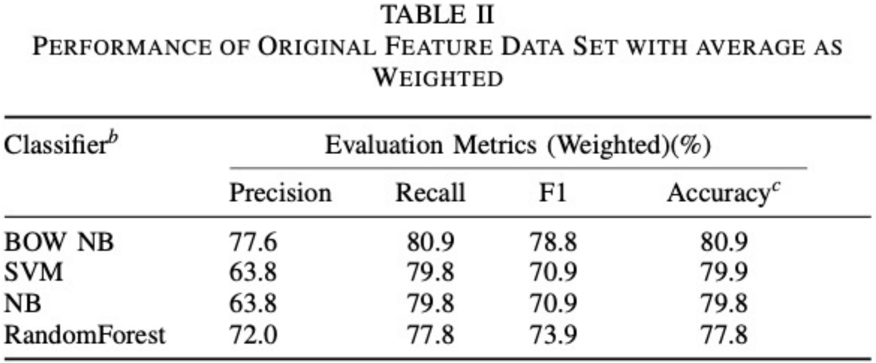
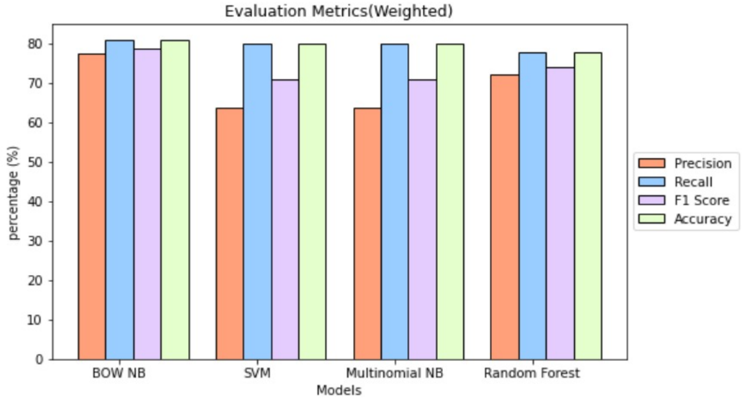
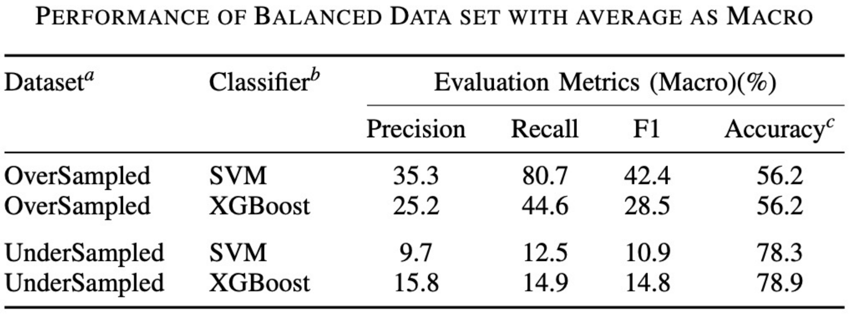


TABLE II
PERFORMANCE OF ORIGINAL FEATURE DATA SET WITH AVERAGE AS WEIGHTED

Classifier[b]	Evaluation Metrics (Weighted)(%)			
	Precision	Recall	F1	Accuracy[c]
BOW NB	77.6	80.9	78.8	80.9
SVM	63.8	79.8	70.9	79.9
NB	63.8	79.8	70.9	79.8
RandomForest	72.0	77.8	73.9	77.8

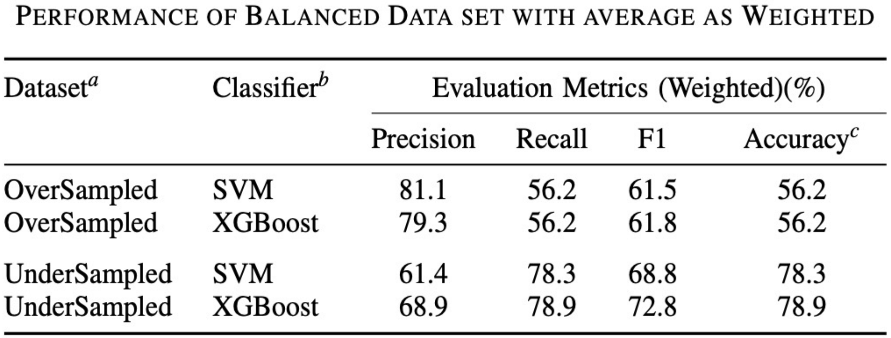


Evaluations - 2 (Balanced Data)



PERFORMANCE OF BALANCED DATA SET WITH AVERAGE AS MACRO

Dataset[a]	Classifier[b]	Evaluation Metrics (Macro)(%)			
		Precision	Recall	F1	Accuracy[c]
OverSampled	SVM	35.3	80.7	42.4	56.2
OverSampled	XGBoost	25.2	44.6	28.5	56.2
UnderSampled	SVM	9.7	12.5	10.9	78.3
UnderSampled	XGBoost	15.8	14.9	14.8	78.9



PERFORMANCE OF BALANCED DATA SET WITH AVERAGE AS WEIGHTED

Dataset[a]	Classifier[b]	Evaluation Metrics (Weighted)(%)			
		Precision	Recall	F1	Accuracy[c]
OverSampled	SVM	81.1	56.2	61.5	56.2
OverSampled	XGBoost	79.3	56.2	61.8	56.2
UnderSampled	SVM	61.4	78.3	68.8	78.3
UnderSampled	XGBoost	68.9	78.9	72.8	78.9

Comparison with BOW:

- Bag-Of-Words (BOW) outperforms selected models considering Accuracy, Precision, Recall values
- BOW suffers from sparse representation and much bigger vocabulary resulting in higher computational complexity

Conclusion:

- Random forest performed really well for this dataset, because of highly imbalanced nature
- Weighted average gave results better than Macro average
- Chunking of the book can be explored to find if it gives better results

Challenges:

- Computation cost in extracting features is high.
- Additional features needs to be explored to enrich the feature set
- Class Imbalance problem