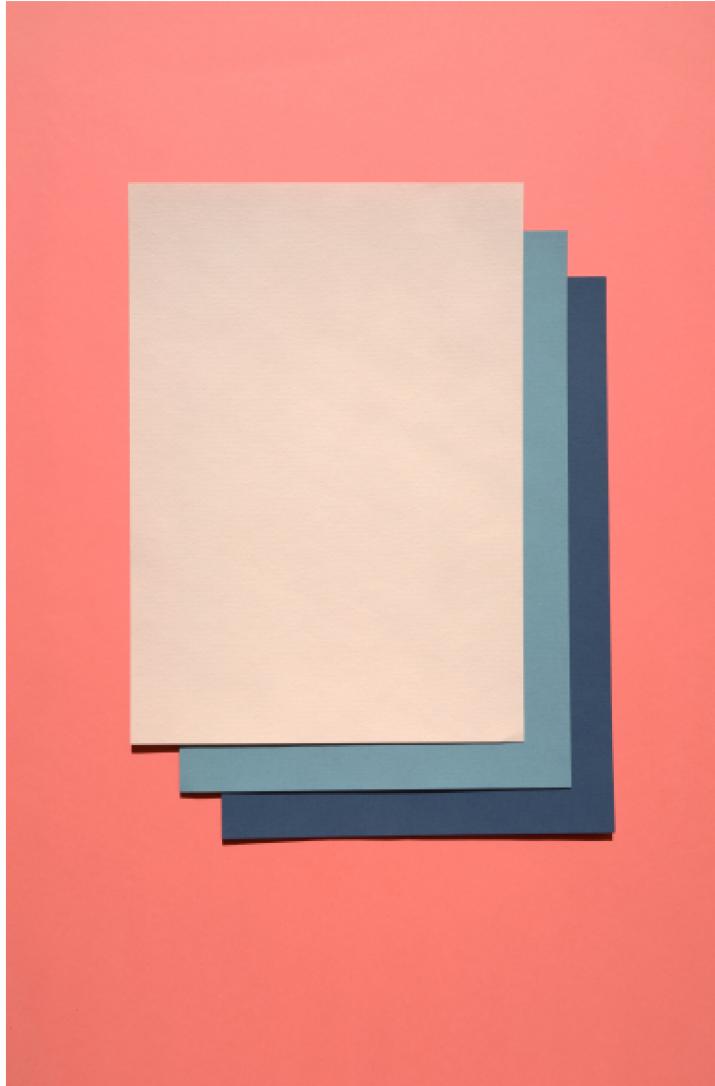


Title	Context	Statement	Methodology	Classes	Testaments	LR vs NB



# NLP Using The Bible

## Mekdes Kebede

# Context

450 translations in English

66 books

40 authors

1500 years



## Problem Statement:

What is the optimal machine learning method to identify a text with similar content from a given corpus?

# Methodology

Number of versions : 6

Classification vs Regression

Data Scient Process

## Processors:

TFIDF

DOC2VEC

## Models:

LR

NB

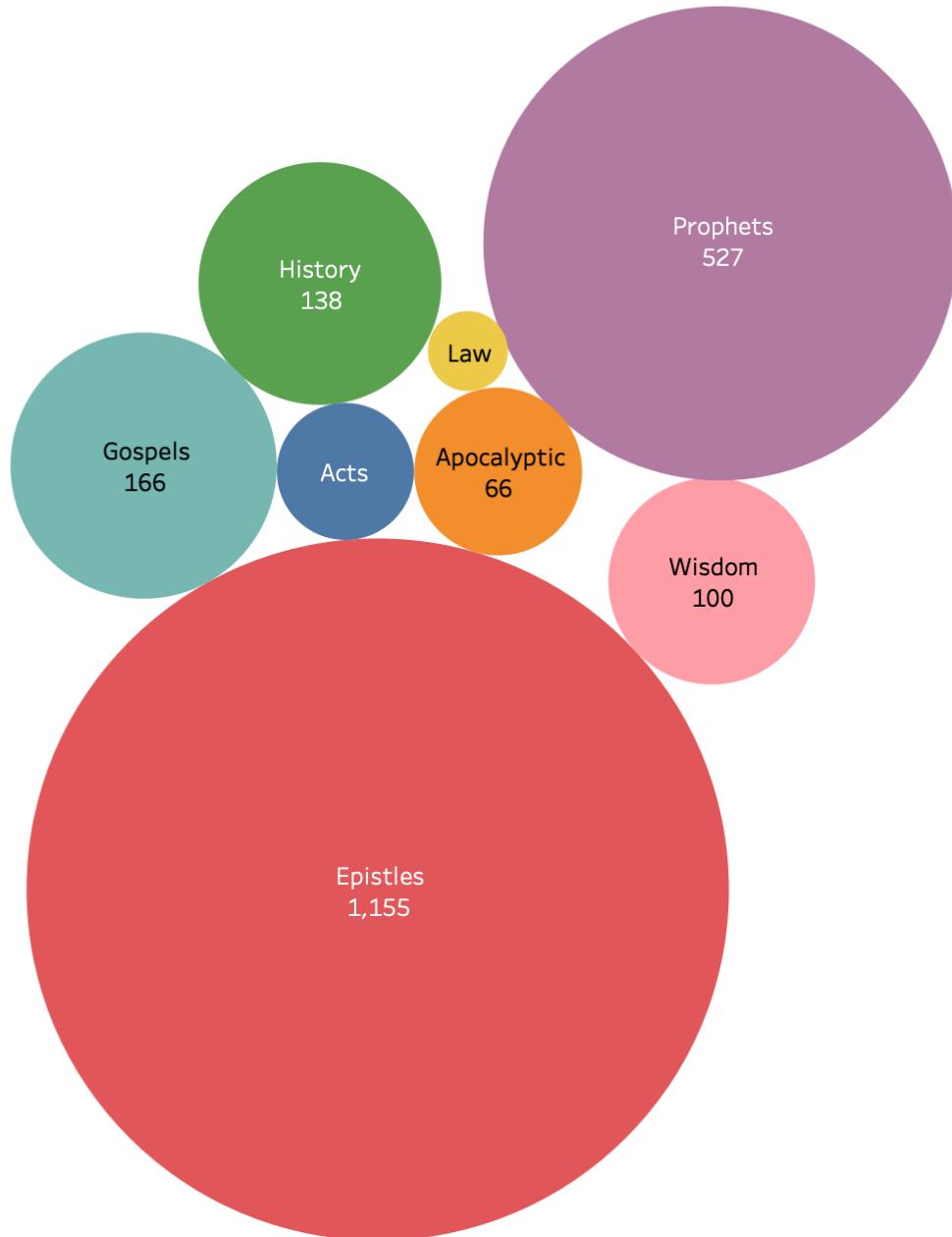
## Evaluation:

Accuracy %

f1 score



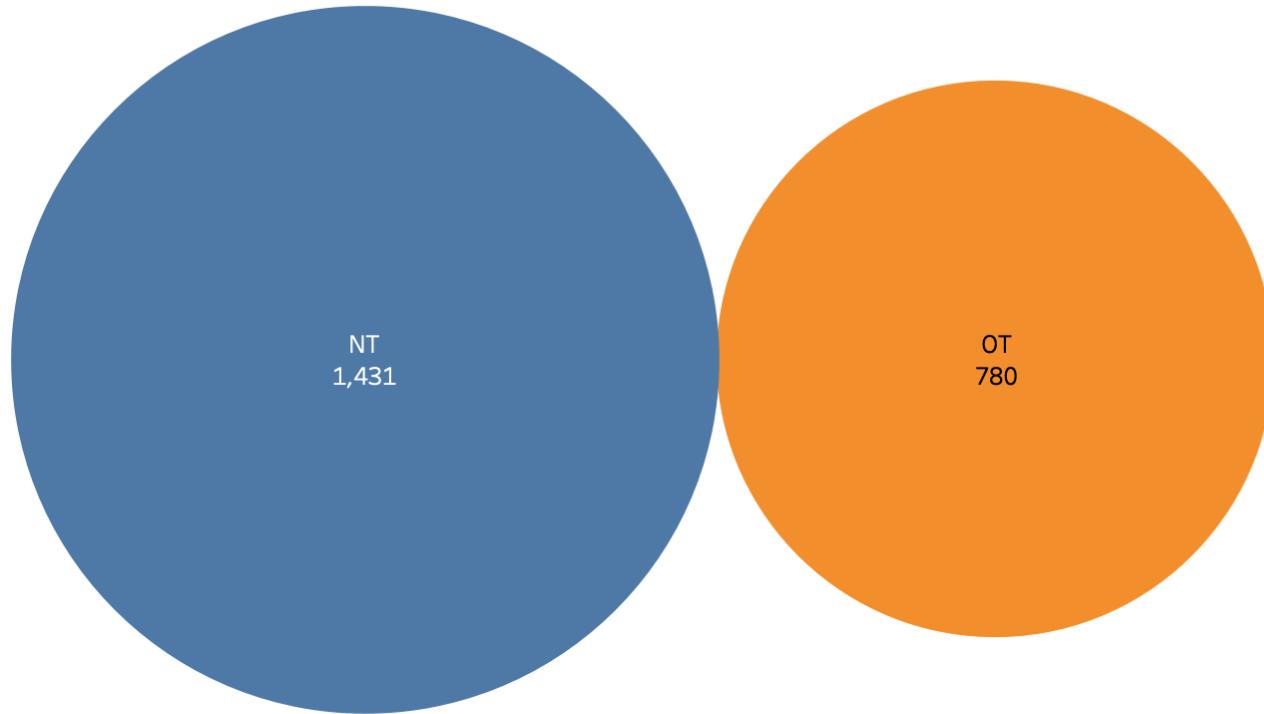
# 8 Genre Classes



## Name of Genres

- Acts
- Apocalyptic
- Epistles
- Gospels
- History
- Law
- Prophets
- Wisdom

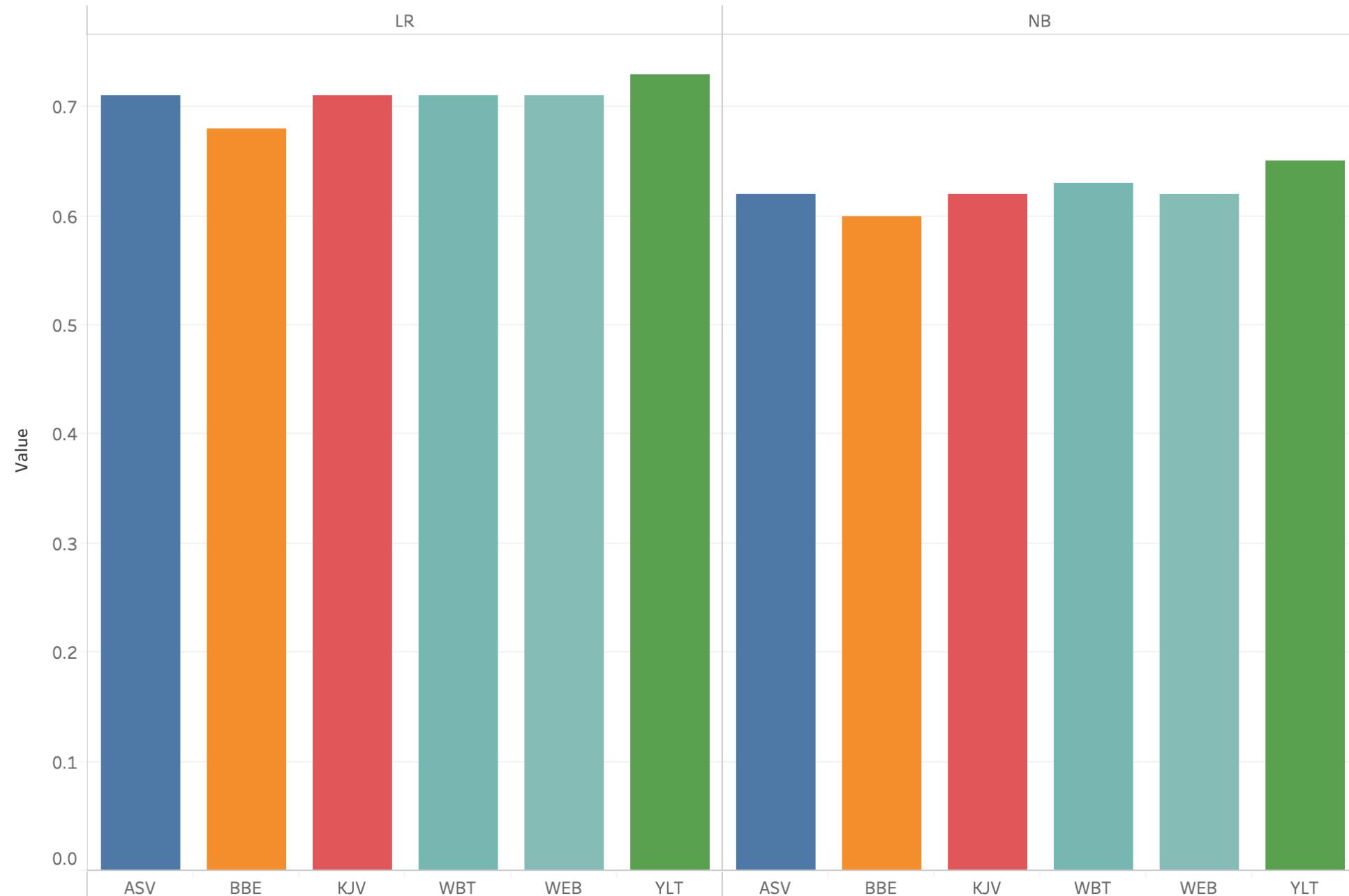
# Old Testament VS. New Testament



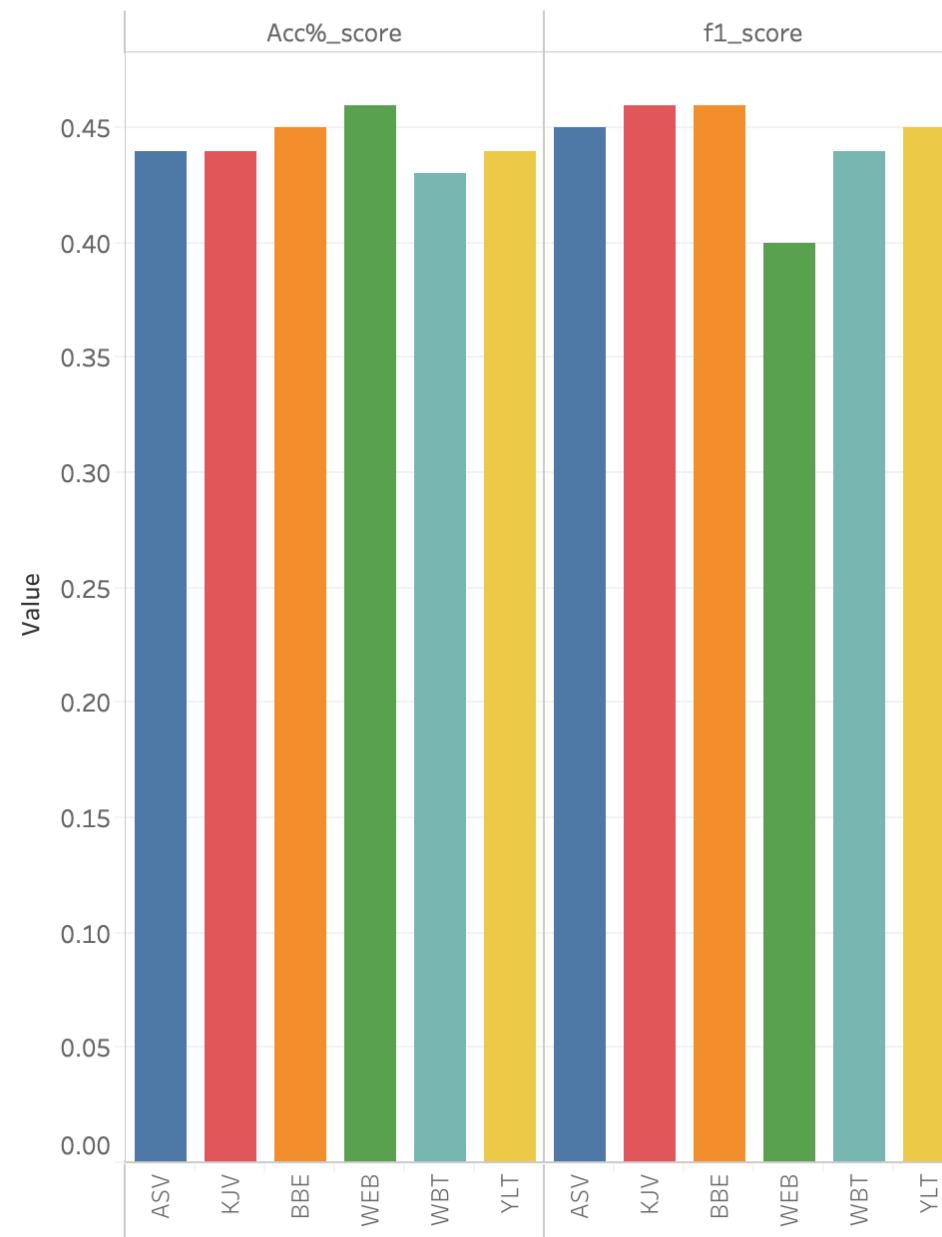
Old VS New Testament

- NT
- OT

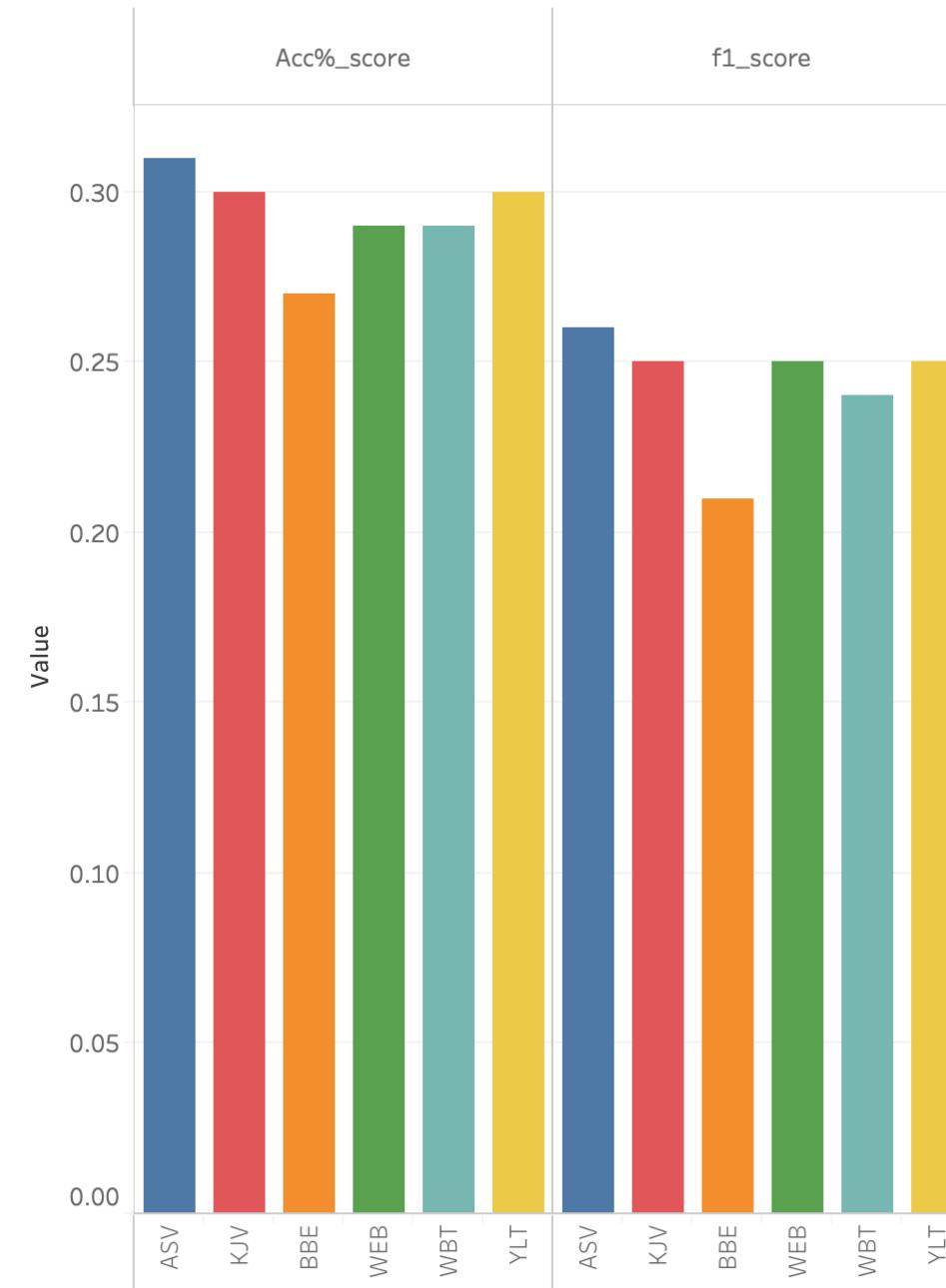
# TFIDF Logistic Regression VS. Naive Bayes

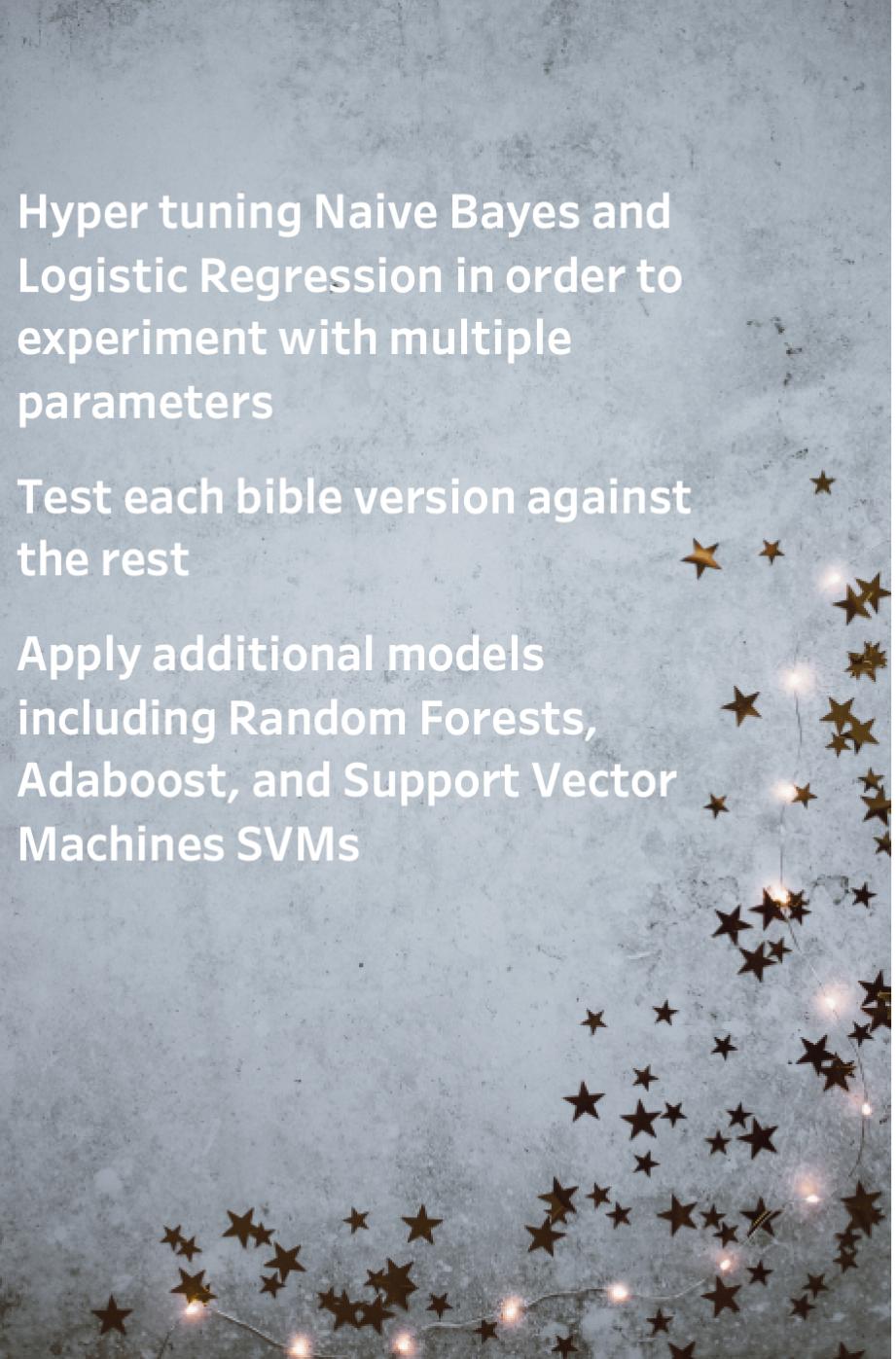


## Distributed Memory(dm=1)



## Distributed Bag of Words(dm=0)





**Hyper tuning Naive Bayes and Logistic Regression in order to experiment with multiple parameters**

**Test each bible version against the rest**

**Apply additional models including Random Forests, Adaboost, and Support Vector Machines SVMs**

**Combine all the versions** and test against a different source of text with the same 8 class genres

Implement a **content-based Recommender** based on the best classification performance above

Utilize a Graphic User Interface (**GUI**) tool to easily be able to input verses in a manner that will recommend similar context in a user-friendly environment

Based on the performance of models and techniques above, **expand the genre keys** as well as versions to train against a new set of genres