Authorship Attribution

Aparna Budhavarapu
1767790

Aashrit Mathur
1611541

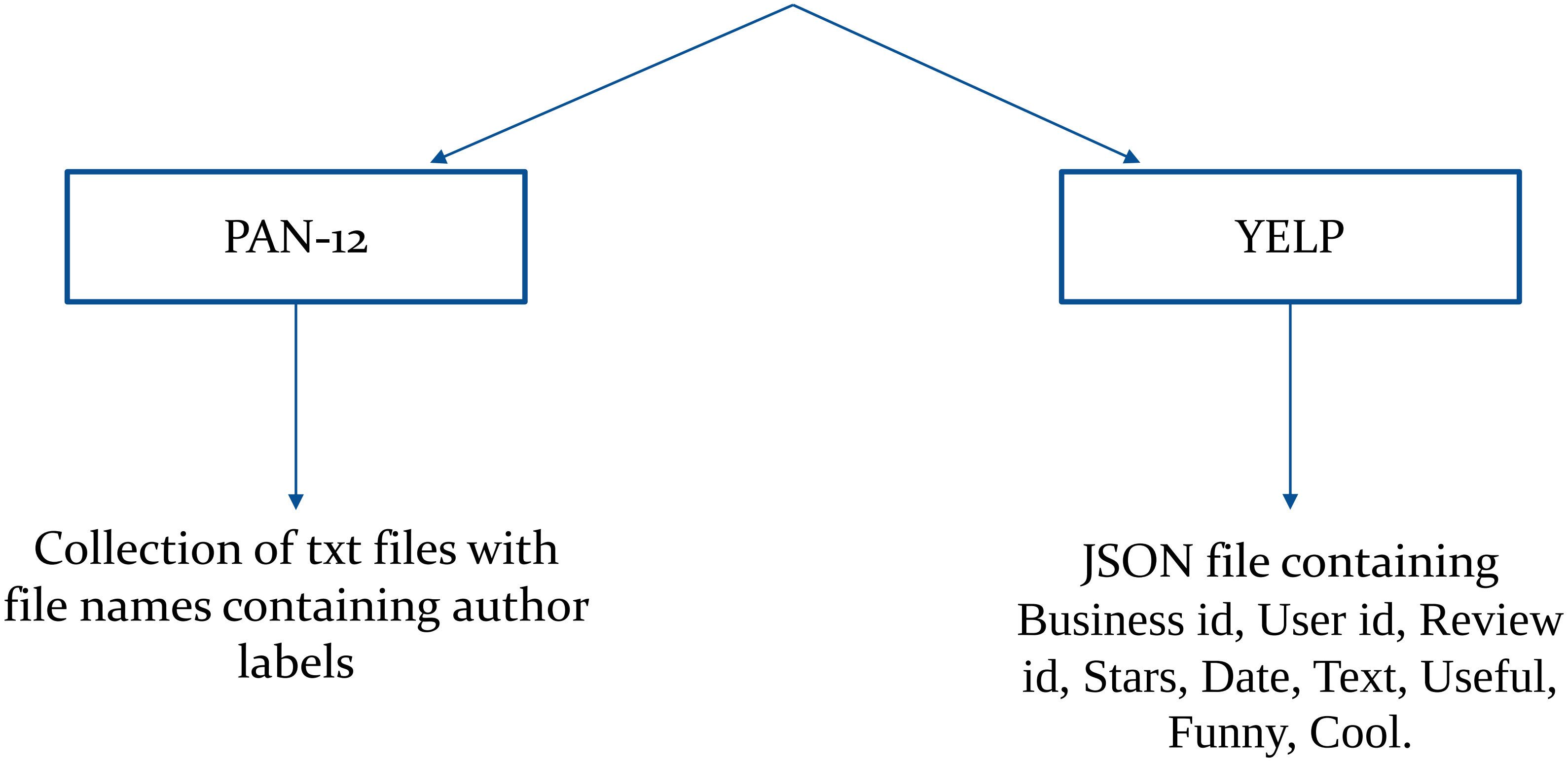
Contents

- Introduction
- Data Used
- Method Implemented
- Performance comparison
- References

Introduction

Given a set of authors, authorship attribution(AA) is the task of figuring out who, if any of them is the actual author of a piece of text.

Data Used



PAN-12

YELP

Collection of txt files with file names containing author labels

JSON file containing Business id, User id, Review id, Stars, Date, Text, Useful, Funny, Cool.

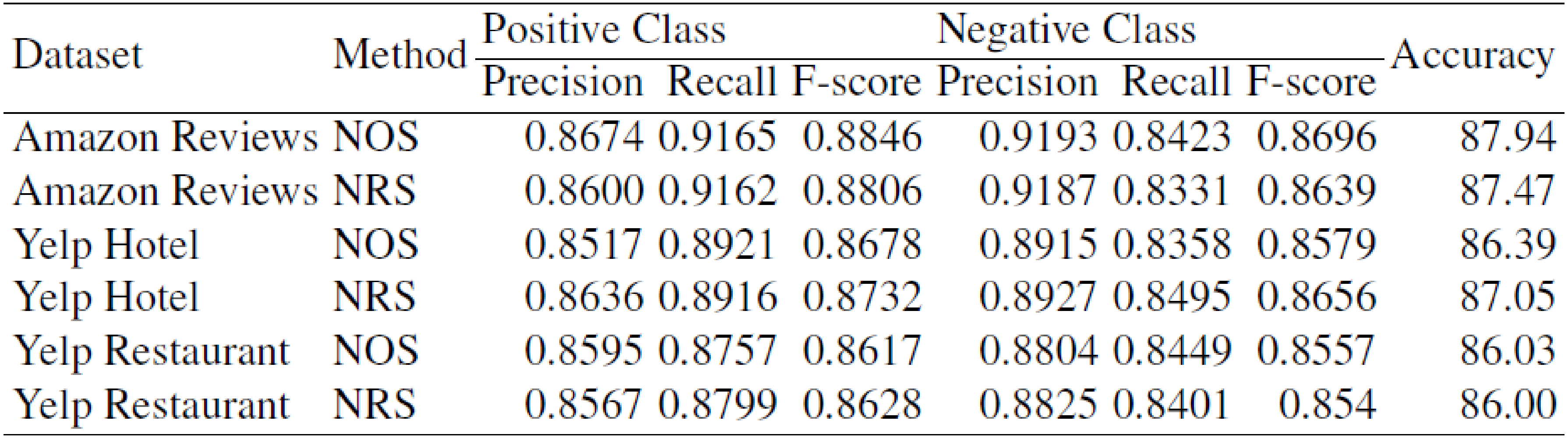
Data used in the baseline paper contains amazon, yelp hotel and yelp restaurant reviews.

Method Implemented

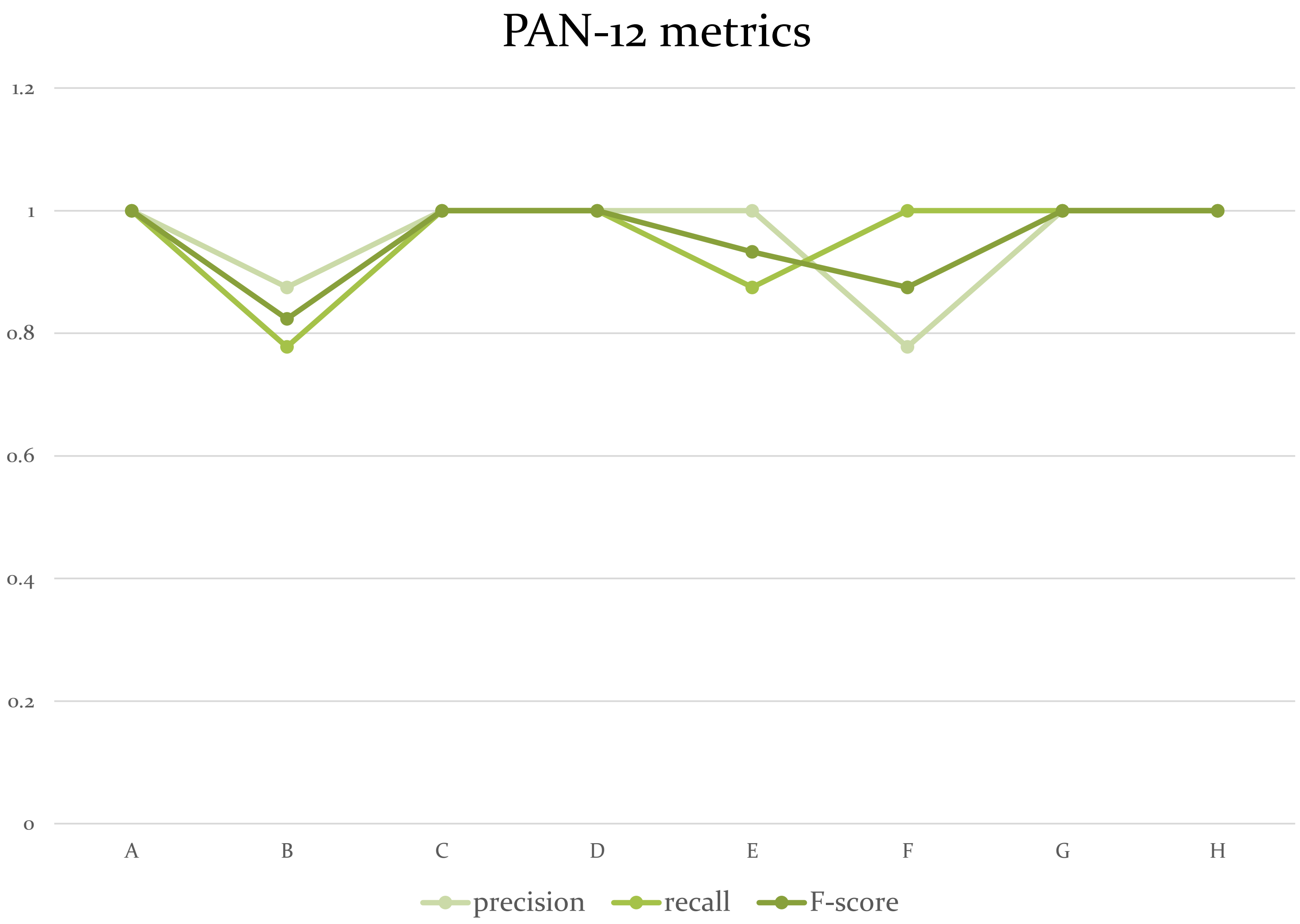
- We took two separate lists:
  - Label
  - Text
- Vectorize the text using TfidVectorizer() for n-grams
- Data split: 80% training and 20% testing
- Build a LinearSVC() model using the training set
- Obtain the accuracy, precision, recall and F-score for the PAN-12 data and accuracies for 1-gram, 2-gram and 3-gram for Yelp data

Performance Metrics

Baseline paper

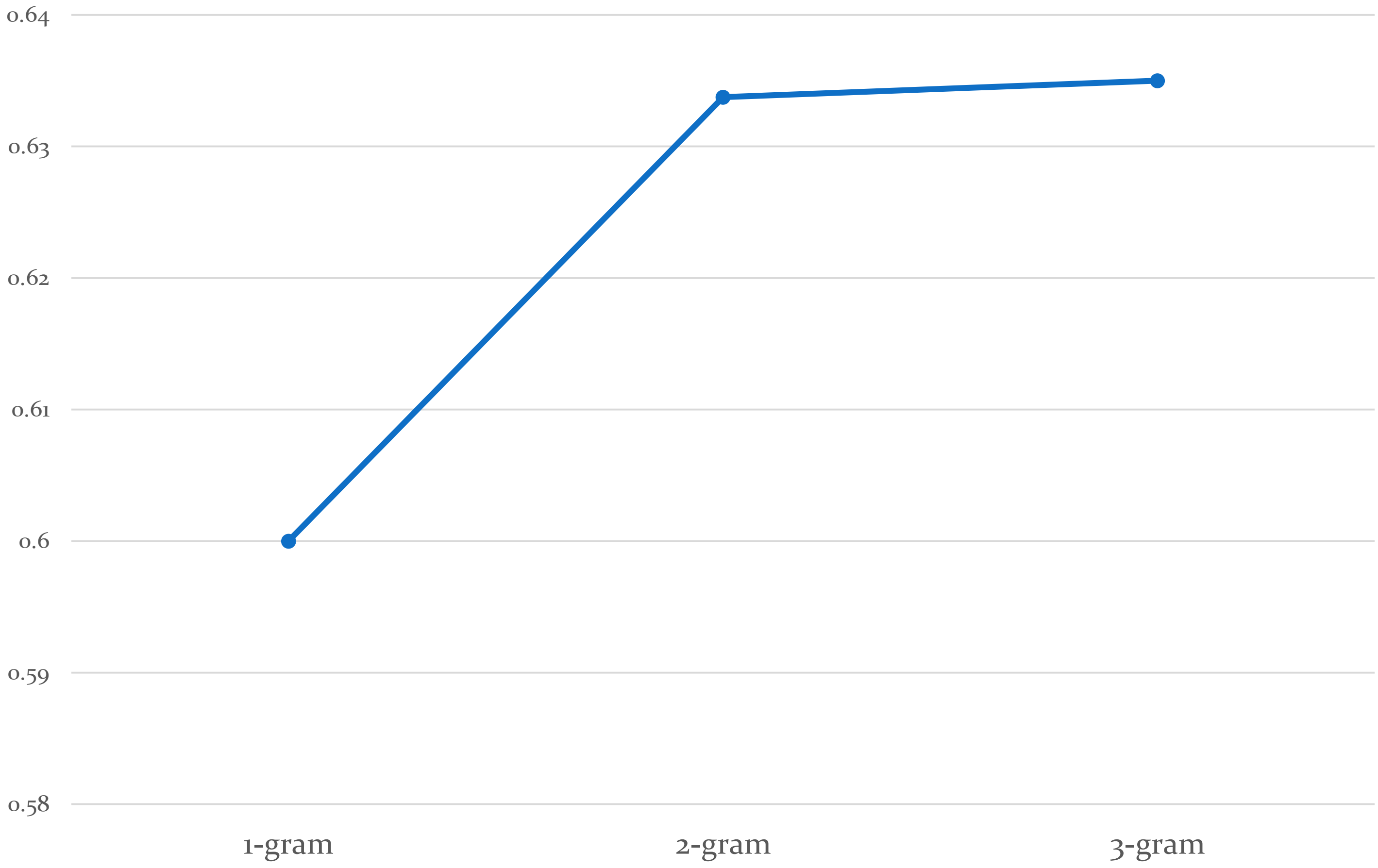


Dataset	Method	Positive Class			Negative Class			Accuracy
		Precision	Recall	F-score	Precision	Recall	F-score	
Amazon Reviews	NOS	0.8674	0.9165	0.8846	0.9193	0.8423	0.8696	87.94
Amazon Reviews	NRS	0.8600	0.9162	0.8806	0.9187	0.8331	0.8639	87.47
Yelp Hotel	NOS	0.8517	0.8921	0.8678	0.8915	0.8358	0.8579	86.39
Yelp Hotel	NRS	0.8636	0.8916	0.8732	0.8927	0.8495	0.8656	87.05
Yelp Restaurant	NOS	0.8595	0.8757	0.8617	0.8804	0.8449	0.8557	86.03
Yelp Restaurant	NRS	0.8567	0.8799	0.8628	0.8825	0.8401	0.854	86.00



Average accuracy obtained = 0.7

Accuracy for Yelp dataset



References

- https://www.yelp.com/dataset/download
- http://www2.cs.uh.edu/~arjun/papers_new/Shrestha%20et%20al.%20CICLING%2016.pdf
- https://blog.michaelckennedy.net/2017/06/21/yelp-reviews-authorship-attribution-with-python-and-scikit-learn/
- https://www.researchgate.net/publication/310799885_Generalized_Confusion_Matrix_for_Multiple_Classes

THANK YOU