Android authorship attribution through analysis of String *n-grams*

Vaibhavi Kalgutkar, Natalia Stakhanova, Paul Cook

Canadian Institute for Cybersecurity - University of New Brunswick

Problem Statement

- Mobile device market, especially Android is expanding rapidly
- Increasing number of malicious apps due to openness of google play store
- Android users are becoming more susceptible to malware
- Need of an automated system to detect such malicious apps
- We propose to develop a lightweight system to generate the signatures for malware writers which in turn will be useful to detect malware samples generated by particular malware author

Methodology

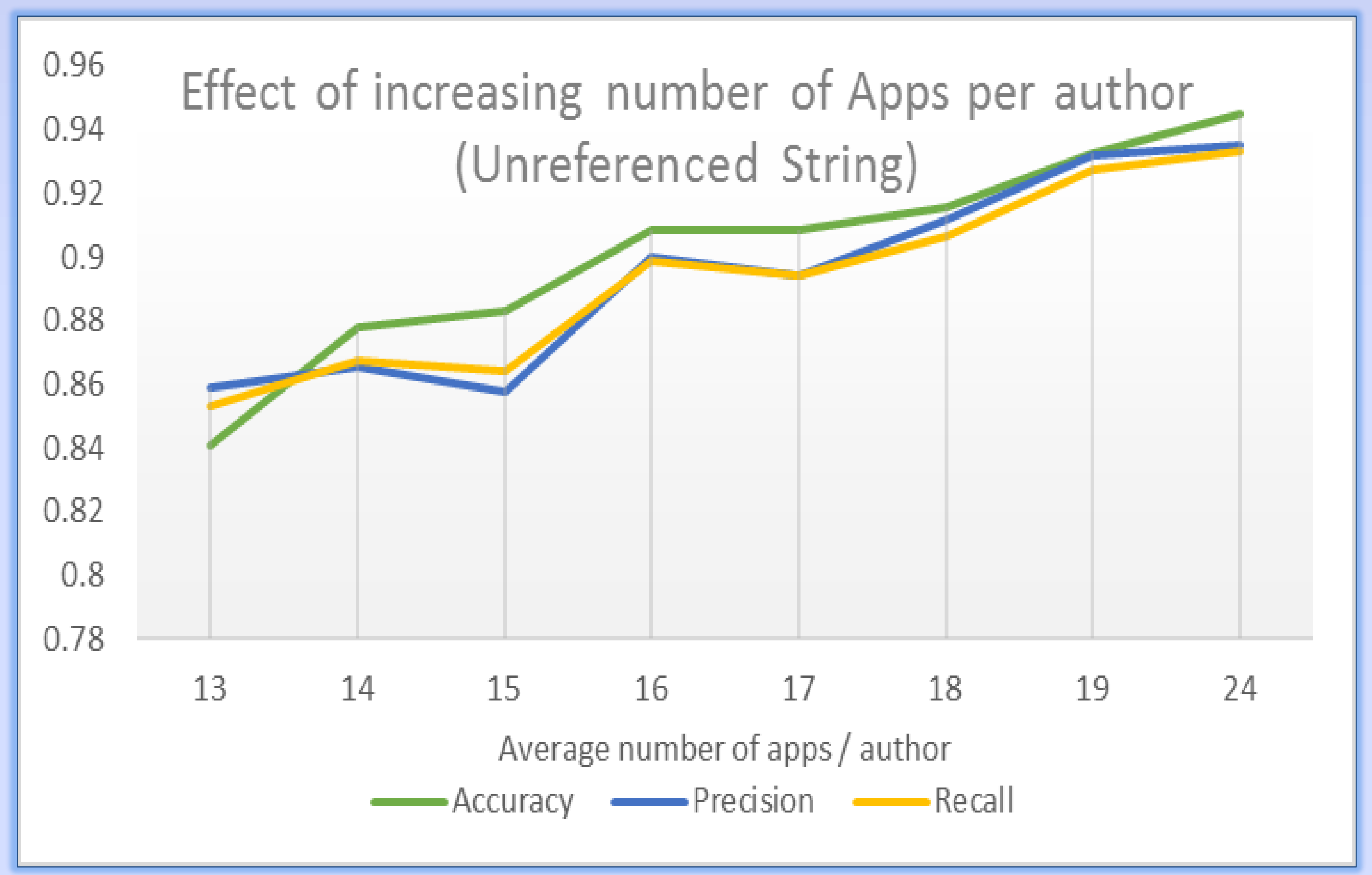
- A machine-learning based approach
- 3-gram word counts are considered
- Three kinds of strings are analyzed namely referenced, unreferenced and application specific strings

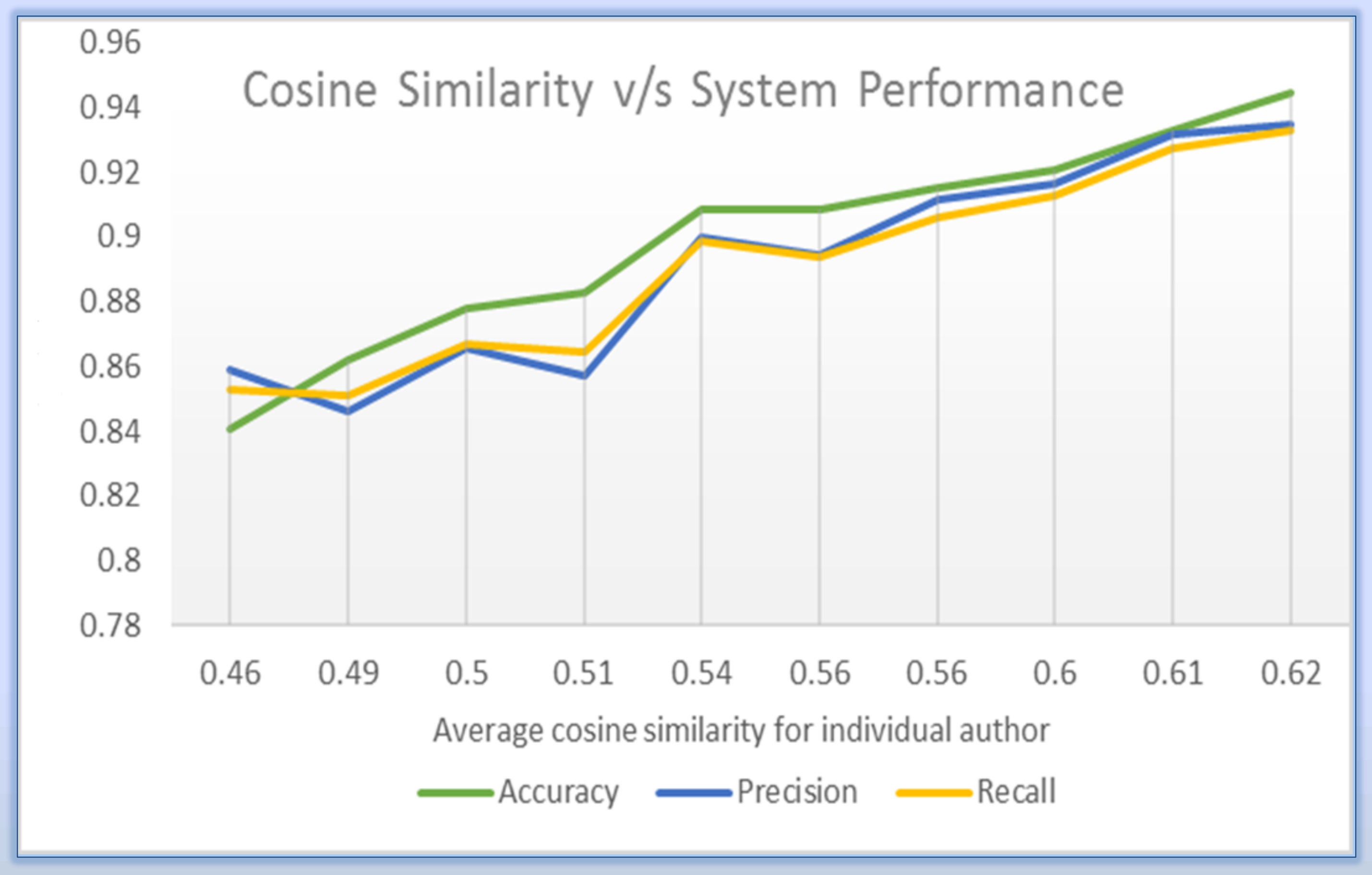
Experimental Setup

- ✓ Benign Dataset : 1599 apps, 40 benign authors
- ✓ Malware Dataset : 266 apps, 10 malicious authors
- ✓ Linear SVM classifier
- ✓ 5 times 5-fold cross validation

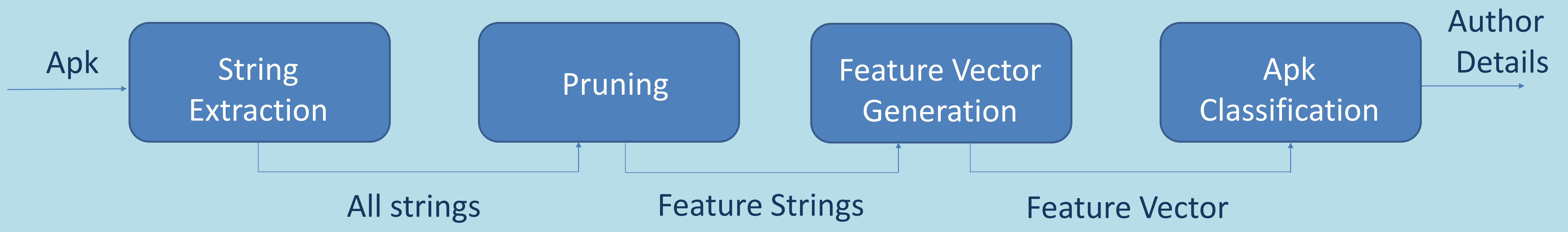
Type of Strings

- **Referenced strings present in DEX file**
  - ✓ Referenced by one of the identifier sections of DEX file
  - ✓ Part of functional app code
- **Unreferenced strings present in DEX file**
  - ✓ Present in the data section of DEX file and only referenced by string offset list
  - ✓ Carry hidden or interesting textual information
- **Strings extracted from strings.xml**
  - ✓ Referenced from the application or from other resource files in APK
  - ✓ Application specific strings defined by the author





System Architecture



Experimental Results

Dataset	String Type	Average Accuracy	Macro Average Precision	Macro Average Recall	Macro Average F1
Benign	Application specific	0.934	0.945	0.919	0.920
Benign	Unreferenced	0.955	0.947	0.939	0.938
Benign	All strings combined	**0.961**	**0.956**	**0.948**	**0.948**
Malicious	Application specific	0.825	0.859	0.848	0.830
Malicious	Unreferenced	0.950	0.959	0.951	0.948
Malicious	All strings combined	**0.962**	**0.971**	**0.964**	**0.962**

Conclusion & Future Work

- We have presented a solution to identify the author of an android app through the use of text strings extracted from the Android Executables file. The proposed system using a Linear SVM with line bounded word level 3-grams was able to identify the authors with an accuracy of 96%
- Future Work : Evaluate performance of the system over set of obfuscated apps