Stratification and categorisation, for South African and global COVID19 social media text data.

Using machine learning and Natural processing methods to extract text features for categorisation, classification and detection.

INTRO

- Using statistical sampling techniques for classifications and micro-blog bots usage detection
- Feature engineering methods are also explored to detect micro-blog location base.

METHODS

1. The data-set used was collected by the team at University of Pretoria and partners.
2. About 1 million data input row and 117 columns data sample was provided.
3. **Feature selection:** TF-IDF vectorization
4. **Classification:** CNN, Logistic regression and Naive beyes classifier.

RESULTS

Location based text classification

- CNN – 0.988
- Logistic regression: 0.990

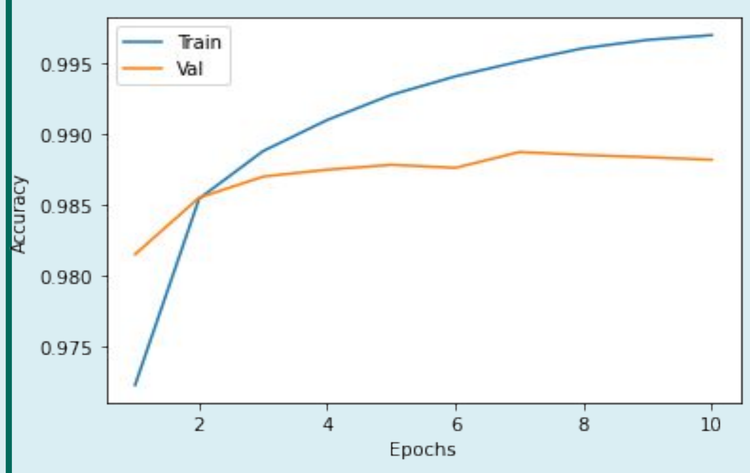
Bots detection and classification

- CNN – 0.872
- Logistic regression- 0.881

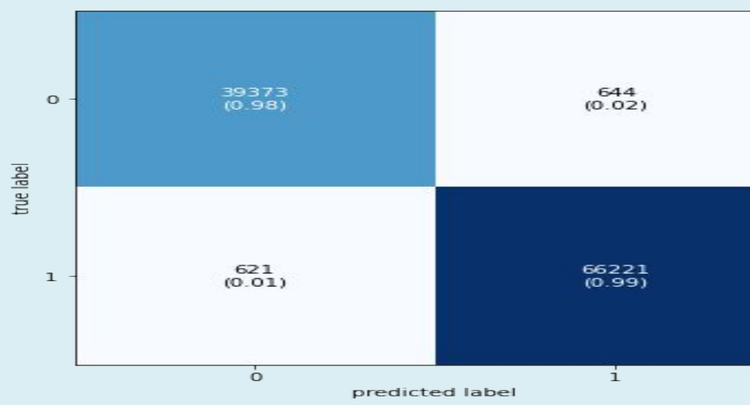
Sentiments analysis

- Logistic regression- 0.930
- CNN – 0.900

CNN model evaluations



Naive bayes classifier evaluations



Nakeng Mohlatlego, Raymond Chiruka

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Department of Computer Science

Faculty of Engineering, Built Environment and Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en Inligtingtegnologie / Lefapha la Boetšenere, Tikologo ya Kago le Theknolotši ya Tshedimošo

Capstone Project - MIT 808

Course Coordinators:
Dr. Vukosi Marivate (vukosi.marivate@cs.up.ac.za)
Abiodun Modupe (abiodun.modupe@cs.up.ac.za)

28 as

Scan me