Evaluating

Logistic regression and random forest models for classifying text formality using abbreviation and their expansion

BY: ABDOUL, BAOWA, KATRINA, MARISOL

CONTENT

INTRO

Analyzing in depth how acronyms and abbreviations are detected and processed in the environments of spelling checkers and autocorrects.

And create a model that can help us understand this process step by step by utilizing logistic regression and random forest.

RESEARCH QUESTION

How effectively can logistic regression and random forest models distinguish between formal and informal texts based on the presence and expansion of abbreviations?

PREVIOUS LITERATURE

- Abbreviation disambiguation is the process of choosing the correct long form of an abbreviation
  - Often based on its context (Song, JIN, Myaeng 2005)
- Dictionary/corpus and machine learning based algorithms are proposed methods in dealing with abbreviation disambiguation (Song, JIN, Myaeng 2005)
- Due to the fact that some abbreviations can represent multiple long form meanings, a dictionary or a corpus may not be wholly satisfying (Henriquez & Hernandez 2009)

METHODS

N° 1

Find a chat corpus (informal and formal) that utilizes acronyms and abbreviations,

However, we were unable to find such corpus.

N° 2

Find a dataset that has various
acronyms and abbreviations.

N° 3

Because of the lack of chat corpus, we created a function that generates a synthetic text by utilizing the acronym/abbreviation datasets.

N° 4

Categorize the text into 5 categories (Acronym, Expansion, Informal text, Tokenized and Formal text.

N° 5

- Data transformation and label assignment

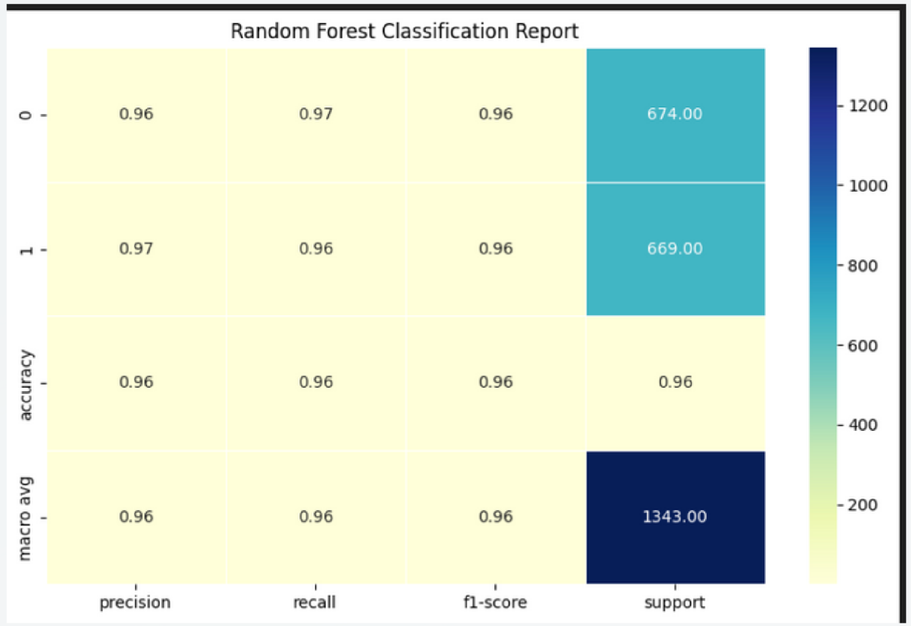
N° 6

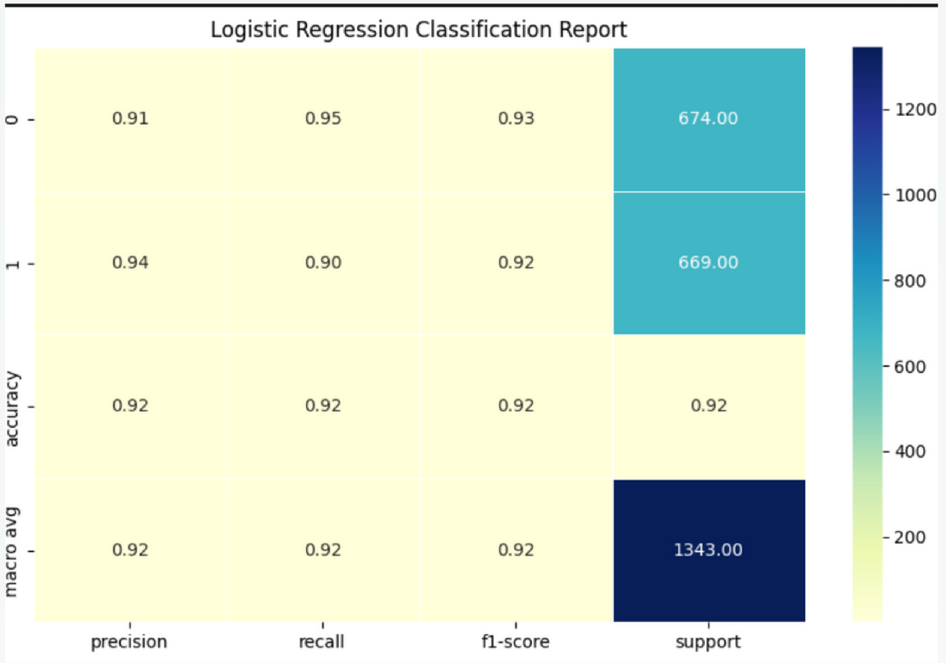
Use logistic regression and random forest to predict and evaluate the accuracy of our dataset.

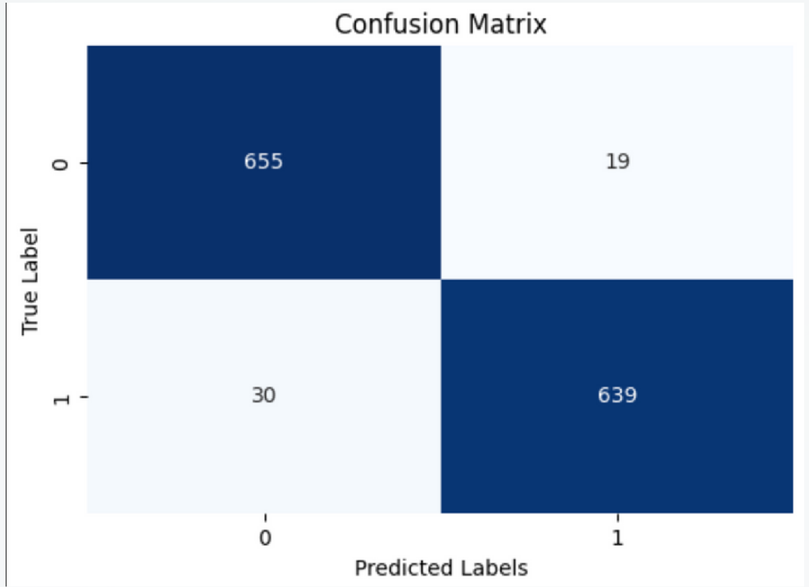
CODE

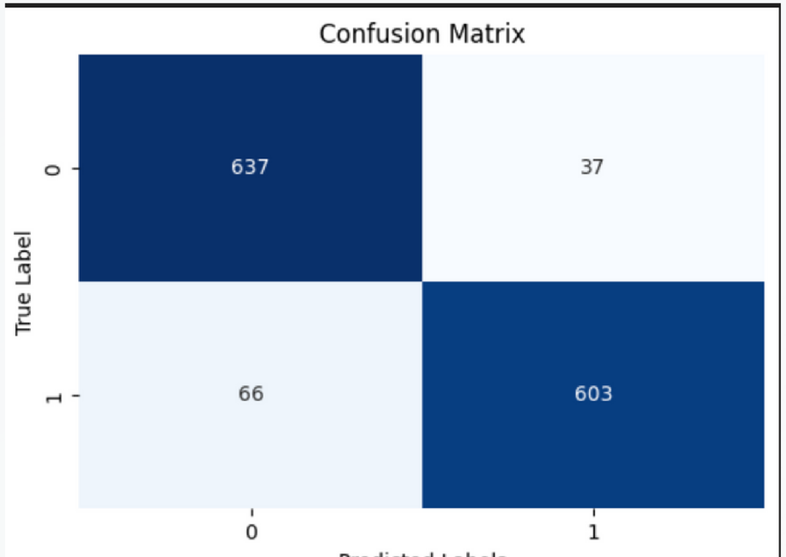
https://github.com/Abdoul1996/chat_slang/blob/main/chat_slang.ipynb

Performance Comparison of Random Forest and Logistic Regression Models



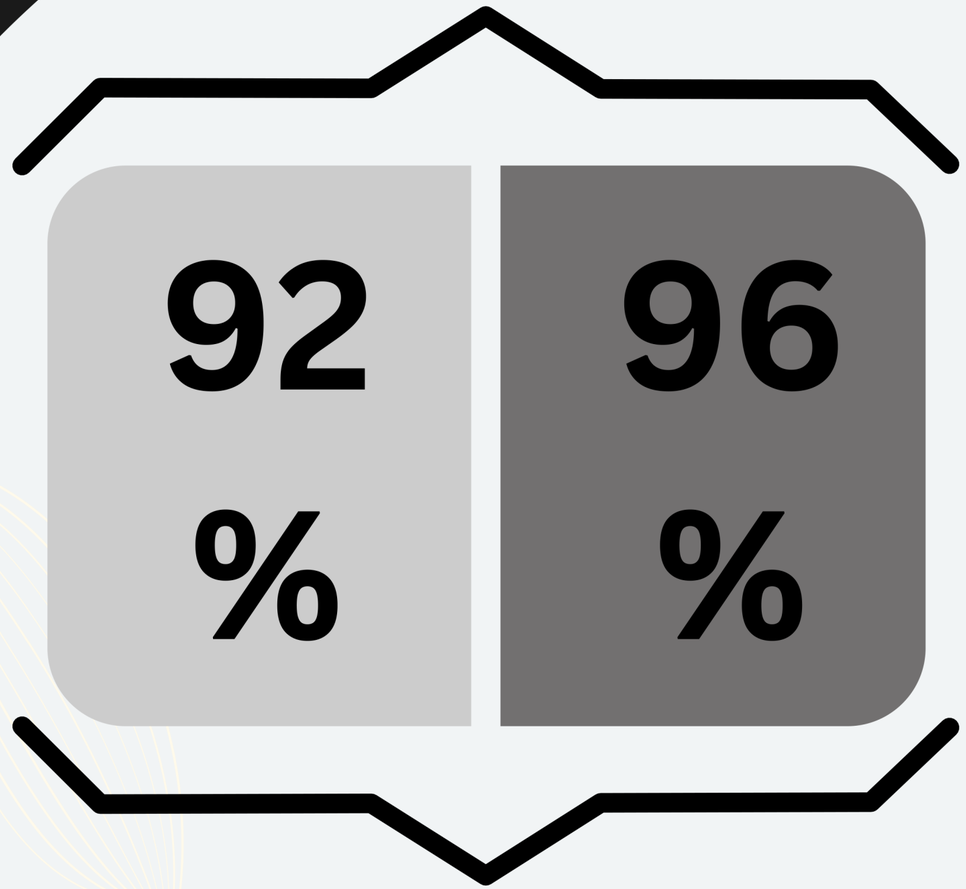






STATISTICS

- The Logisitic Regression model yielded 92% accuracy
- The RandomForest model yielded 96% accuracy



Summary

- Initially, we planned to study how modern acronyms and abbreviations affect the accuracy of spelling error detection. However, due to lack of suitable data and complexities we have modified our objective of the project.
- To overcome, we developed a function that generates a synthetic text by randomly incorporating acronyms into sentences, thus creating a artificial samples of text.

- We then trained two machine learning models and found that the Random Forest model identifies acronyms with greater accuracy than Logistic Regression, with performance scores of 96% and 92%
- Finally, this project demonstrate an innovative approach to creating a dataset when existing data does not meet research needs, highlighting the use of synthetic text as a viable solution for specific linguistic analyses.

FURTHER RESEARCH

Create a function that can check sentence structure

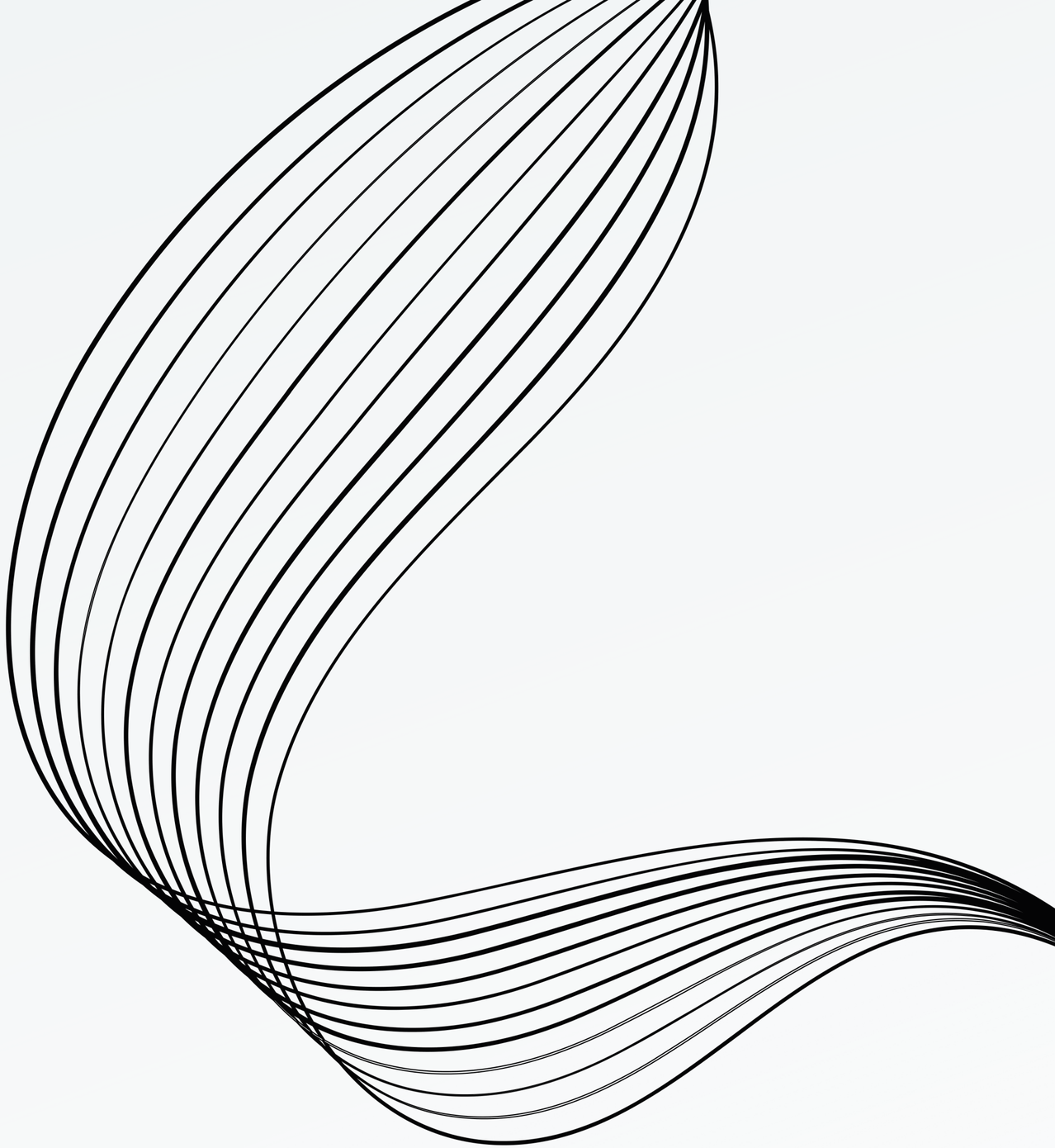
Double check that every single letter is not being processed as an acronym or abbreviation

Both models can be improved to get higher Accuracy

Formal Grammars and Parsing

Using different hyperparameter tuninng to each model could affect their performance

QUESTION?



THANK YOU!