HATE SPEECH DETECTION IN KENYA

1 Computers possess have cognitive abilities to understand the human language the way we do unless we train them to .This project aimed at aimed at training computer models to identify/classify hate speech on twitter(Location :\*\*KENYA\*\*)

Therefore, this project aims to automate content moderation to identify hate speech using machine learning techniques in Kenya.

Kenya is a multilingual country [with over 42 tribes];

Kenyans tweet in English & Swahili [Others mix their tweets with a bit of Vernacular]. This also occurs when people are conversating. This is what is known as Codeswitching (Speaker alternates between two or more languages when either speaking or in this case \*\*TWEETING\*\*)

\*\*DEFINATIONS OF HATE SPEECH\*\*

Hate speech is defined in a variety of ways in the literature, and there is \*\*no universal definition\*\*. Other organizations, such as international and domestic legislation, have attempted to define hate speech and even identified specific targets based on what are legally known as protected traits, such as race, ethnic origin, religion, or gender.

Here are a few definitions of what hate \*\*speech entails\*\*

\*\*[Twitter hateful conduct policy]\*\*: “You may not promote violence against or directly attack or threaten other people on the basis of race, ethnicity, national origin, sexual orientation, gender, gender identity, religious affiliation, age, disability, or disease. We also do not allow accounts whose primary purpose is inciting harm towards others on the basis of these categories.”

\*\*[Kenya National Cohesion and Integration Commission (NCIC) Act, 2008]\*\*: “Content that promotes violence against or has the primary purpose of inciting hatred against individuals or groups based on certain attributes, such as race or ethnic origin, religion, disability, gender, age, veteran status, sexual orientation/gender identity.”

\*\*[The European Court of Human Rights] \*\*: “All forms of expression which spread, incite, promote or justify racial hatred, xenophobia, anti-Semitism or other forms of hatred based on intolerance, including intolerance expressed by aggressive nationalism and ethnocentrism, discrimination and hostility towards minorities, migrants and people of immigrant origin.”

\*\*Hate speech should not be conflicted with freedom of speech\*\*

There is freedom of speech for what to speak/air out their opinions, but these should not give one a room to be offensive or use hate speech.

Baseline models included Random Forest, Naive Bayes, Logistic Regression and Support Vector Machine (SVM). The final model was a \*\*Logistic Regression\*\* model that used Count Vectorization for feature engineering. It produced an F1 of 0.3958 and Recall (TPR) of 0.624.

Check out the project repository [here](http