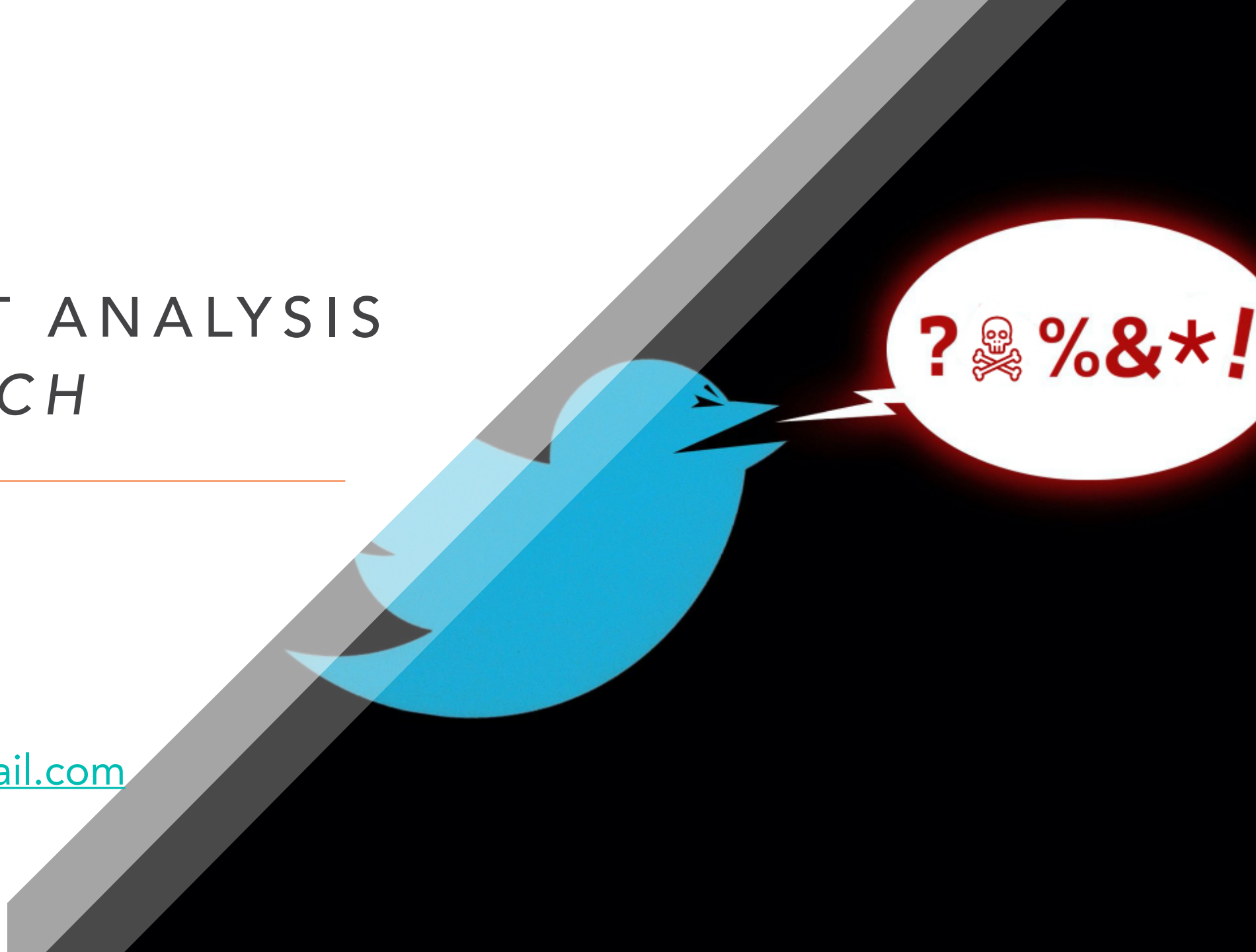
TWITTER SENTIMENT ANALYSIS *HATE SPEECH*



Matteo Tortella

Capstone Project

Linkedin Page

matteotortella4@gmail.com

Hate Speech on Twitter

Twitter's conduct policy has recently been under the scrutiny of many as the social platform is unfortunately thought to be the largest forum for diffusion of dehumanising speech.

One of the possible reasons behind this, it's the underlying anonymity which especially characterises it. On one hand, not having a real identity on the platform has its advantages when it comes to speak out in oppressive regimes. On the other, it also promotes **hate speech.**

On top of relatively ineffective guidelines or rules, another major problem for the platform is the sheer amount of tweets that are generated per second.

The main goal of this project is to build a dummy pipeline to help Twitter identify and block hateful tweets in a **proactive and almost real-time basis.**

PROJECT WORKFLOW AND METHODOLOGY

This project is divided into three main sections :

1) Assembled 40,000 human labelled tweets (20000 positive and 20000 negative) in order to train Machine Learning models classify offensive tweets from good ones.

2) Performed an iterative optimisation process on selected models across Neural Networks, Logistic Regression and Naive Bayes. The current best performing (Logistic Regression) has a 98.3% accuracy rate.

3) In the final section, our winning model is ran across freshly tweets that are stored on a local MySQL database. The model predictions are then analysed and streamed on an ***online app*** showcasing number of offensive tweets along with their text.

40,000

TWEETS ANALYSED ON DATASET

98%

ACCURACY ON TEST SET

61

AVERAGE NUMBER OF CHARACTERS IN OFFENSIVE TWEETS

The next slide contains content which is offensive and objectionable

Limitations and Future Work

- OUT-OF-WORDS VOCABULARY
- LIMITED TO ENGLISH LANGUAGE
- INHERENT BIAS IN THE HUMAN LABELLING OF TWEETS

Samples of WordClouds produced in the project