# TWITTER SENTIMENT

Speaker

Mohamed Emran

# Introduction

## Problem..

- The problem is to analyze user behavior on Twitter, by using Natural Language Processing (NLP), to determine if the tweet is either positive or negative.

## Value..

- By analyzing user-behavior, we gain a deep understanding towards specific topics, products or brands
- Businesses can also monitor user in real-time, by responding to any issues before it escalates.

# Data Collection

## Where did I get my data?

## The Source..

Data obtained from a research group in Stanford University (http://help.sentiment140.com/for-students/)

## Collected From..

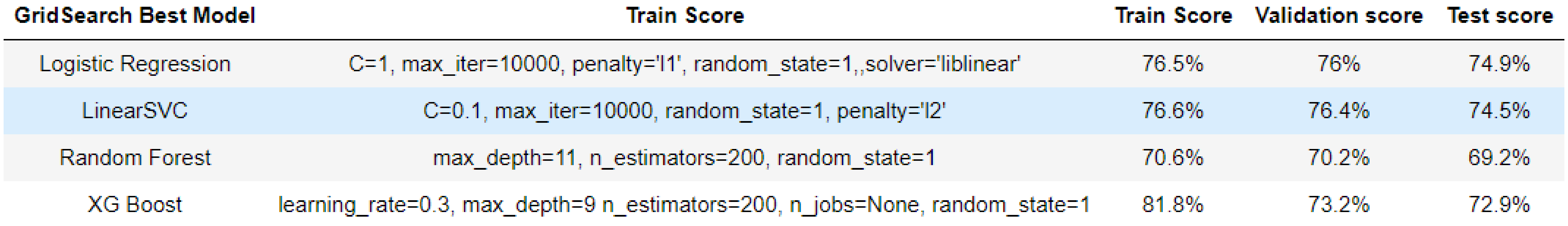
Kaggle: **Sentiment140 dataset with 1.6 million tweets**

# Preprocessing

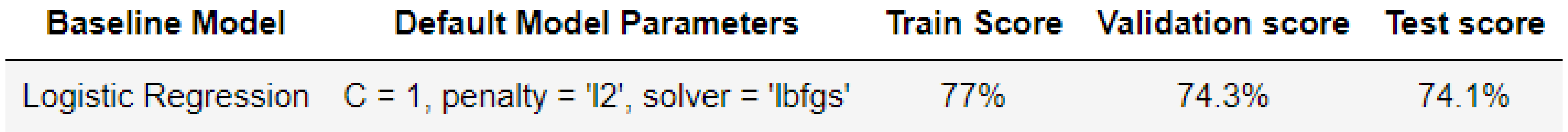
1. Sampling dataset
   - 10% ~ 160k
2. Vectorizing
   - Using CountVectorizer
   - I tested out stemmer and lemmatizer

# Modeling

1. Logisitic Regression
2. SVM
3. Decision Trees
4. XG Boost



| GridSearch Best Model | Train Score | Train Score | Validation score | Test score |
|---|---|---|---|---|
| Logistic Regression | C=1, max_iter=10000, penalty='l1', random_state=1,,solver='liblinear' | 76.5% | 76% | 74.9% |
| LinearSVC | C=0.1, max_iter=10000, random_state=1, penalty='l2' | 76.6% | 76.4% | 74.5% |
| Random Forest | max_depth=11, n_estimators=200, random_state=1 | 70.6% | 70.2% | 69.2% |
| XG Boost | learning_rate=0.3, max_depth=9 n_estimators=200, n_jobs=None, random_state=1 | 81.8% | 73.2% | 72.9% |



| Baseline Model | Default Model Parameters | Train Score | Validation score | Test score |
|---|---|---|---|---|
| Logistic Regression | C = 1, penalty = 'l2', solver = 'lbfgs' | 77% | 74.3% | 74.1% |

# Thank You!