Predicting Clash Royale Tweets

- A JSC270 Project

By Steven Liu, William Zhang

Problem Description and Motivation

Question: Can we build a model that predicts whether a tweet is in favor of the recent game updates, against them, or neutral (stating fact / no opinion)?

Difficulties:

- Usually a mix of opinions
- Need instant feedback
- Lots of sarcasm, internet-slang, & abbr

Currently not many similar analysis

Motivation

- Thinking from the developer’s pov
- Great tool to “collect” feedback (vs traditional survey way)

Related analysis:

1. Twitter sentiment analysis of game reviews (on a much larger scale) : classify attitude towards a game

Description of Data

- Entries: 397 Tweets searched with keyword

  “@ClashRoyale Update” and “@ClashRoyale Balance”

- Self Labeled Outcomes:

  0 = neutral, 1=in favor, 2=not in favor

- 70-30 Split for Train-Test
- Restricted to Tweets only
- Set max of 1000 features (feature vector)

Description continued

Strength: Self-Labeled, Strong opinions, Recent and Relevant

Limitation: Human inconsistency, collected within short period of time, contain repeated tweets (promotions & etc.), ambiguous wording

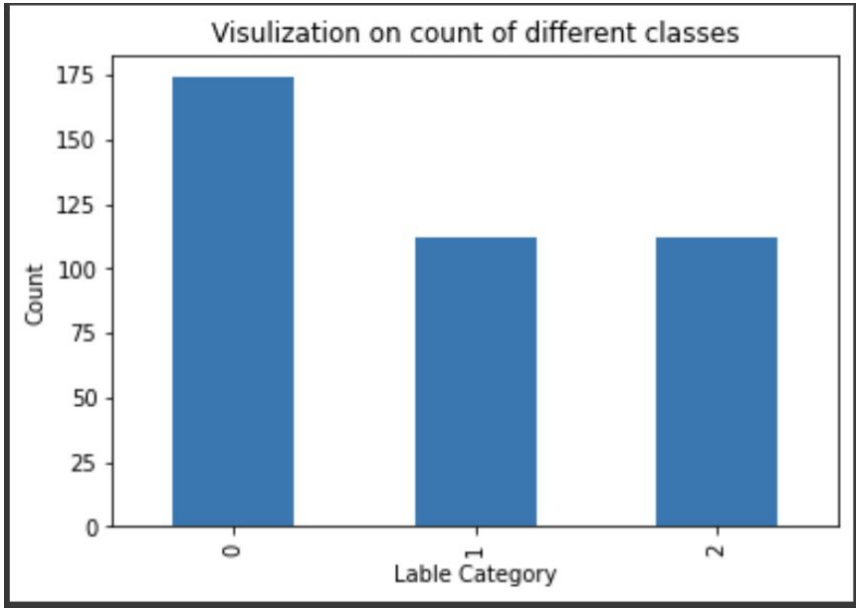
Ex: Can i get a kitten to help me cope with this big new update?

- Twitter analysis of game reviews: 21000 tweets to train, took very similar approaches (in terms of EDA and modeling), 64-83-90 accuracy (Bayes, SVM, Max Entropy)

Exploratory Data Analysis

Standard NLP Procedures

- Data Cleaning
- Tokenization
- Lemmatization
- Remove stop words
- Remove punctuations and special characters (a lot come from unreadable emojis)
- CountVectorizer & TFIDF & Classification Tree



Majority being neutral, equal share of like and dislike

Machine Learning Model

- Classification Tree & Naive Bayes & Support Vector Classifier (SVC)

Classification Tree

- Train Accuracy: 0.583（58%）
- Test Accuracy: 0.508（51%）

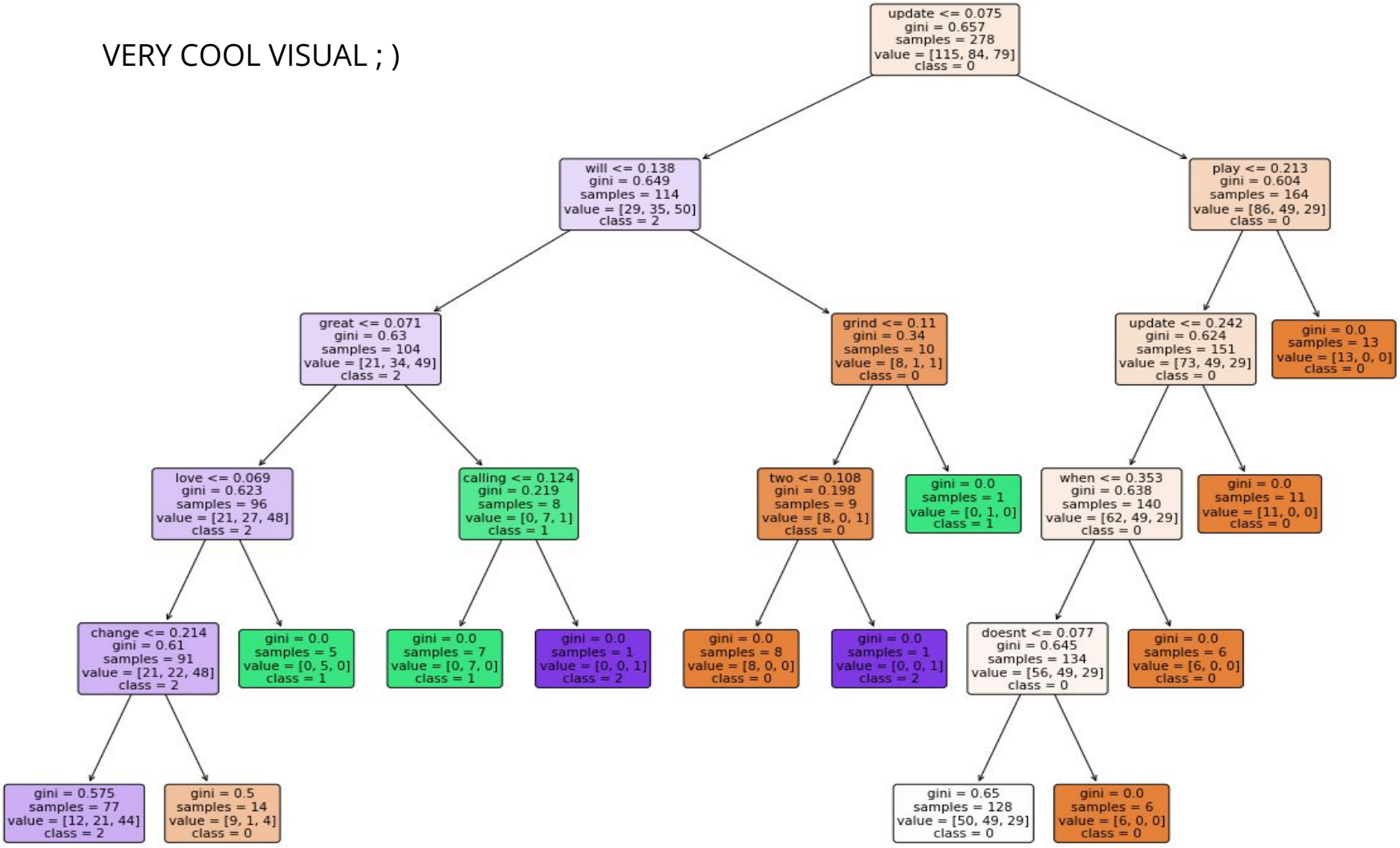
Confusion Matrix: (on test data)

	Predicted Neutral (0)	Predicted Positive (1)	Predicted Negative (2)
Actual Neutral	43	0	16
Actual Positive	20	3	5
Actual Negative	18	0	15

Test Accuracy on Class 1 & 2: 0.295 (29.5%)

Horrible Performance...
But..

VERY COOL VISUAL ;)



Naive Bayes

- Supervised Learning

Using Countvectorizer:

- Train Accuracy: 0.924（92%）
- Test Accuracy: 0.666（67%）

Confusion Matrix: (on test data)

	Predicted Neutral (0)	Predicted Positive (1)	Predicted Negative (2)
Actual Neutral	43	8	8
Actual Positive	8	20	0
Actual Negative	9	7	17

Test Accuracy on Class 1 & 2: 0.606 (61%)

Using TFIDF:

- Train Accuracy: 0.917 (92%)
- Test Accuracy: 0.666 (67%)

Confusion Matrix:

	Predicted Neutral (0)	Predicted Positive (1)	Predicted Negative (2)
Actual Neutral	53	4	2
Actual Positive	12	16	0
Actual Negative	17	5	11

Test Accuracy on Class 1 & 2: 0.442 (44%)

Support Vector Classifier

- Supervised Learning

- Train Accuracy: 0.968（97%）
- Test Accuracy: 0.650（65%）

Confusion Matrix: (on test data)

	Predicted Neutral (0)	Predicted Positive (1)	Predicted Negative (2)
Actual Neutral	44	7	8
Actual Positive	7	20	1
Actual Negative	13	6	14

Test Accuracy on Class 1 & 2: 0.557 (56%)

Results and Conclusion

Model Performance:

- Naive Bayes + CountVectorization >= SVC > Naive Bayes + TFIDF > Classification Decision Tree
- Overfitting -> Size of DataSet is small
- Better than random
- Though very Efficient

Improvements:

- More data
- Smarter EDA Process-> Removing repeated tweets

Citation and References

1. Kiran, T. D. V. et al. “Twitter sentiment analysis of game reviews using machine learning techniques.” (2016).

Thank you!