

A blue parallelogram and a light green parallelogram are positioned on the left side of the slide, overlapping each other and the dark background. The blue shape is on the left, and the green shape is to its right, partially overlapping it.


Filtering noise from the NHL Twitterverse

Troy Yang



Motivation

- Twitter generates a lot of content
- A lot of noise
- A lot of bland, informational tweets
- Only interested in “interesting” tweets
- Is there a way to capture this information in a less manual way?



What is not “interesting”?

- Good morning to absolutely no one but the playoff bound Boston Bruins.
- The Mayor of Walpole @chriswags23 ties the game at 2 #NHLBruins
- Can confirm Lee Stempniak has signed a one way contract, \$650,000 with the #nhlbruins
- Personalised Boston Bruins Jersey Throw Pillow Cushion Cover F by RepublicMarketGoods



What is “interesting”?

- Bruins defenseman Connor Clifton, born in Long Branch, N.J., and raised in Matawan, is expected to be in the lineup vs the Devils. Will mark his first NHL game against New Jersey. Should have a nice following tonight.
- #NHLBruins had been getting away with some emerging sloppiness in their puck-mgmt, but not vs. a desperate, recent 2-time champ. Just as well it happens now because the B's need to clean it up to stay on the rails. Grizz is obviously a concern.



Theory

- Tweets containing sentiment, positive or negative, are much more likely to be “interesting” than ones that contain no sentiment



Opening Stages

- Grabbed 100 tweets and some engagement metrics about the Boston Bruins
- Manually label tweets
- Look for a pre-existing solution to establish a baseline



VADER

- Valence Aware Dictionary and sEntiment Reasoner
- Specifically trained for social media posts
- Generalizable to multiple domains
- No training
- Agreed with my labels on 73 out of the 100 tweets I labelled



Pre-processing

- Remove usernames, hashtags, URLs, emojis, excess punctuation and excess whitespace
- Lemmatization



Beating VADER

- Baseline established, could it be beat?
- Sklearn
- Feature extraction
 - BOW, TF-IDF
- Models
 - Logistic Regression, Naive Bayes, Adaboost
- Dataset
 - Airline Sentiment dataset from Kaggle



Beating VADER - Results

- Tuned parameters using 3 fold cross validation on 70% of the data
- Remaining 30% used as testing set to compare models against each other
- VADER did worse than any of the 3 tuned models



Beating VADER - Results

- Logistic Regression

Class	F1	Precision	Recall	Support
Neutral	0.580081	0.633824	0.534739	806.0
Positive	0.672582	0.713748	0.635906	596.0
Negative	0.857871	0.824418	0.894154	2258.0

- VADER

Class	F1	Precision	Recall	Support
Neutral	0.402649	0.385205	0.421749	3099.0
Positive	0.472926	0.325965	0.861193	2363.0
Negative	0.630800	0.893885	0.487361	9178.0



Beating VADER - Results

- Naive Bayes

Class	F1	Precision	Recall	Support
Neutral	0.550642	0.647651	0.478908	806.0
Positive	0.656557	0.700952	0.617450	596.0
Negative	0.857619	0.810161	0.910983	2258.0

- Adaboost

Class	F1	Precision	Recall	Support
Neutral	0.571956	0.567073	0.576923	806.0
Positive	0.641343	0.677239	0.609060	596.0
Negative	0.842613	0.834201	0.851196	2258.0



Beating VADER - Results

- Evaluated the tuned logistic regression model on the 100 tweets I labeled
- Agreed with 77 out of the 100 tweets
- Four more than Vader!



Validation/Outcome

- Grabbed 100 new tweets
- Asked two people to label them
- Compare their labels with tuned logistic regression labels
- Out of the 100, the two validators agreed on 87 of the tweets, and of those 87 my model agreed with them on 68 of them (~80%)
- Sentiment seems to be a “not perfect but definitely better than nothing” way to filter tweets



Areas for future work

- Acquire a more robust dataset in order to great a more generalizable model
- Capture data for the season for different teams so that sentiment can be tracked for different teams
- Compare against team performance, intuitively one would expect better team performance to indicate more positive sentiment
- Use more sophisticated models/feature extraction techniques



Thanks!