Tweet Classification

Daniel Stiner, Brandon Maxwell, Curtis Ullerich

Problem

Classify a tweet as about a company or not.

Ashley staples an apple. ”AHHHH!!!! Apple juice in my eye!”

New Sharp display for Apple? http://t.co/5kvCmfpp

Problem

Twitter presents interesting NLP problems

@Apple please make a squid #emoji so I can put one next to @squidneyy22's name

*Nvm i'll buy you an ipad ^_^*

Problem

Cool ranch , 4 berry sundae , apple juice #yessuh #latenight #snack http://t.co/ttc3vujp

I hate Siri.. -___-

Problem

#Apple's #Retina revolutionhttp://t.co/3PA1TtRb@Imaginalab_Intl

Our Approach

Gather corpus from Twitter with keywords

apple, itunes, ipad, ipod, mac, ios, iphone, AAPL, cupertino, safari, ilife, iwork, garageband, ibook, powerbook, itouch, app store, macworld, facetime, icloud, mobileme, siri, imovie, iphoto, quicktime, logic pro, think different

Our Approach

Using Naive Bayes

- Baseline of 74% accuracy, 81% precision
- Attempt to improve this using preprocessing and better tokenization

Our Approach

@NieveenTunkar an apple , cucumber , tomato , natural valley bars , low fat yogurt or the flavored one .. R so e ideas for snacks !

Stop word removal

@NieveenTunkar apple cucumber tomato natural valley bars low fat yogurt flavored ideas snacks

Our Approach

I hda to get cute today. Apple bottom jeans and fur boots today. #sheratchet

Bigrams

i hda i_hda to hda_to get to_get cute get_cute today cute_today Apple today_Apple bottom Apple_bottom jeans bottom_jeans and jeans_and fur and_fur boots fur_boots today boots_today #sheratchet

Our Approach

@Khrizteenah it's a green apple. It's sour as #$^%^& (>.<) I like it tho

HTML Corrections

@Khrizteenah it's a green apple. It's sour as #$^%^& (>.<) I like it tho

Our Approach

https://t.co/Nin1z9wA my girls

URL replacement

iTunes - Music - I'll Stand By You (The X Factor USA Performance) - Single by Fifth Harmony my girls

Our Approach

Note to self: engraving "My Sweetheart Nancy" on an iPad, makes it more difficult to sell.

Stemming

note to self engrav my sweetheart nanci on an ipad make it more difficult to sell

Our Approach

#Apple’s #Retina revolutionhttp://t.co/3PA1TtRb@Imaginalab_Int

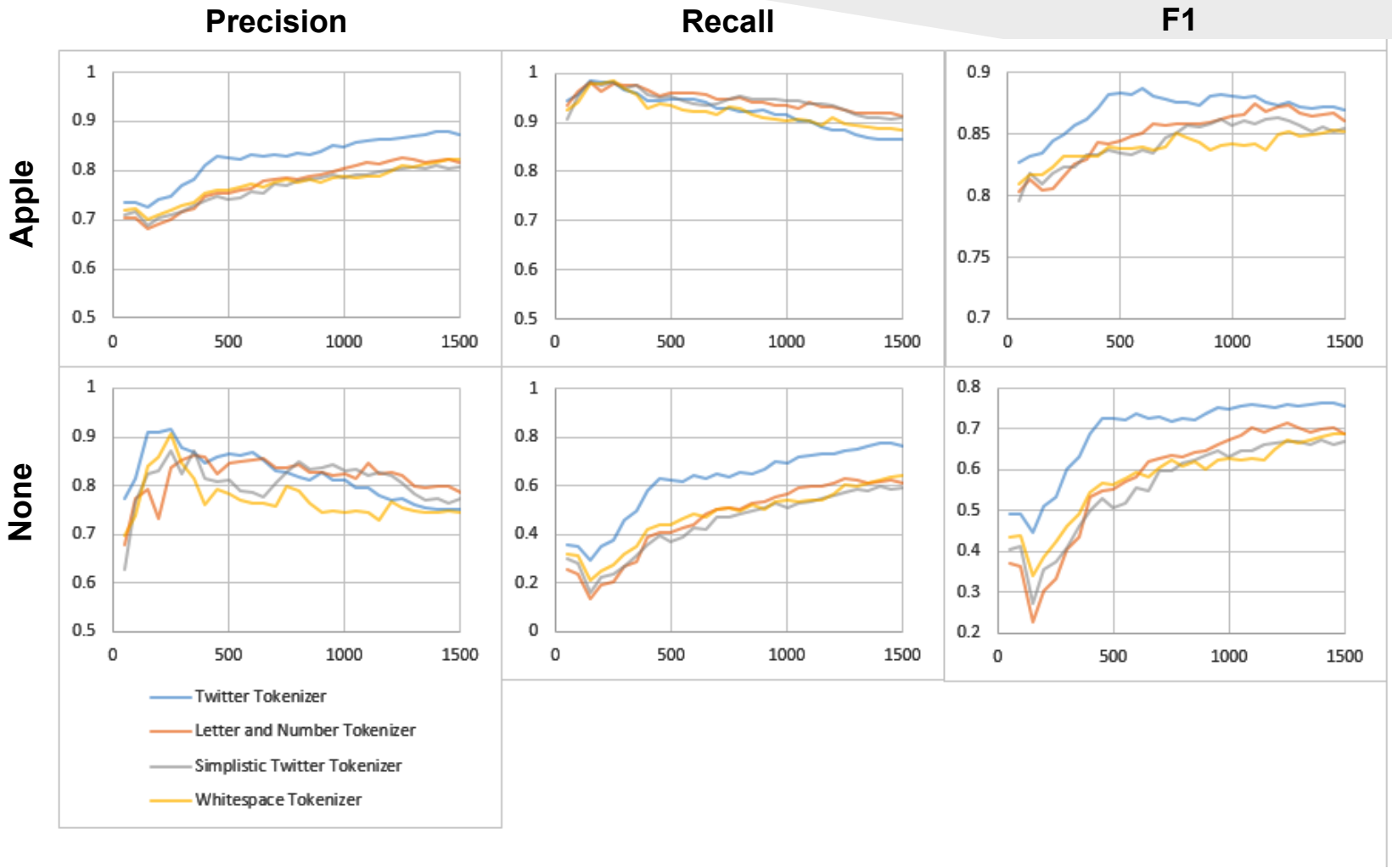
Twitter-aware tokenization

http://t.co/3PA1TtRb @imaginalab_intl #apple #retina s revolution

Results

- Small, but cumulative effect from each step
- Twitter-aware tokenization with Stemming, Bigrams, Stopword Removal, etc.
  - 92.2% precision
  - 84.4% accuracy
- Improvements
  - 12% over whitespace tokenization
  - 13% over alphanumeric tokenization

Results


Precision
Recall
F1
Apple
None
Twitter Tokenizer
Letter and Number Tokenizer
Simplistic Twitter Tokenizer
Whitespace Tokenizer

Contributions

Extensions to Mallet framework
  Pipes and testing

Layered tokenization approach

Questions?

Visit tweets.curtisullerich.com for our paper and code.