Phone Sentiment Analysis with Web Crawl Data

Tuomo Kareoja

Alert! Analytics

November 1, 2019

Agenda

Summary of Findings
- People Like iPhone More
- . . . but the Training Data is Sketchy
- . . . and the Websites Mentioning iPhone are Weird

What We Did?
- Crawl Websites and Count Word Instances
- Modelling

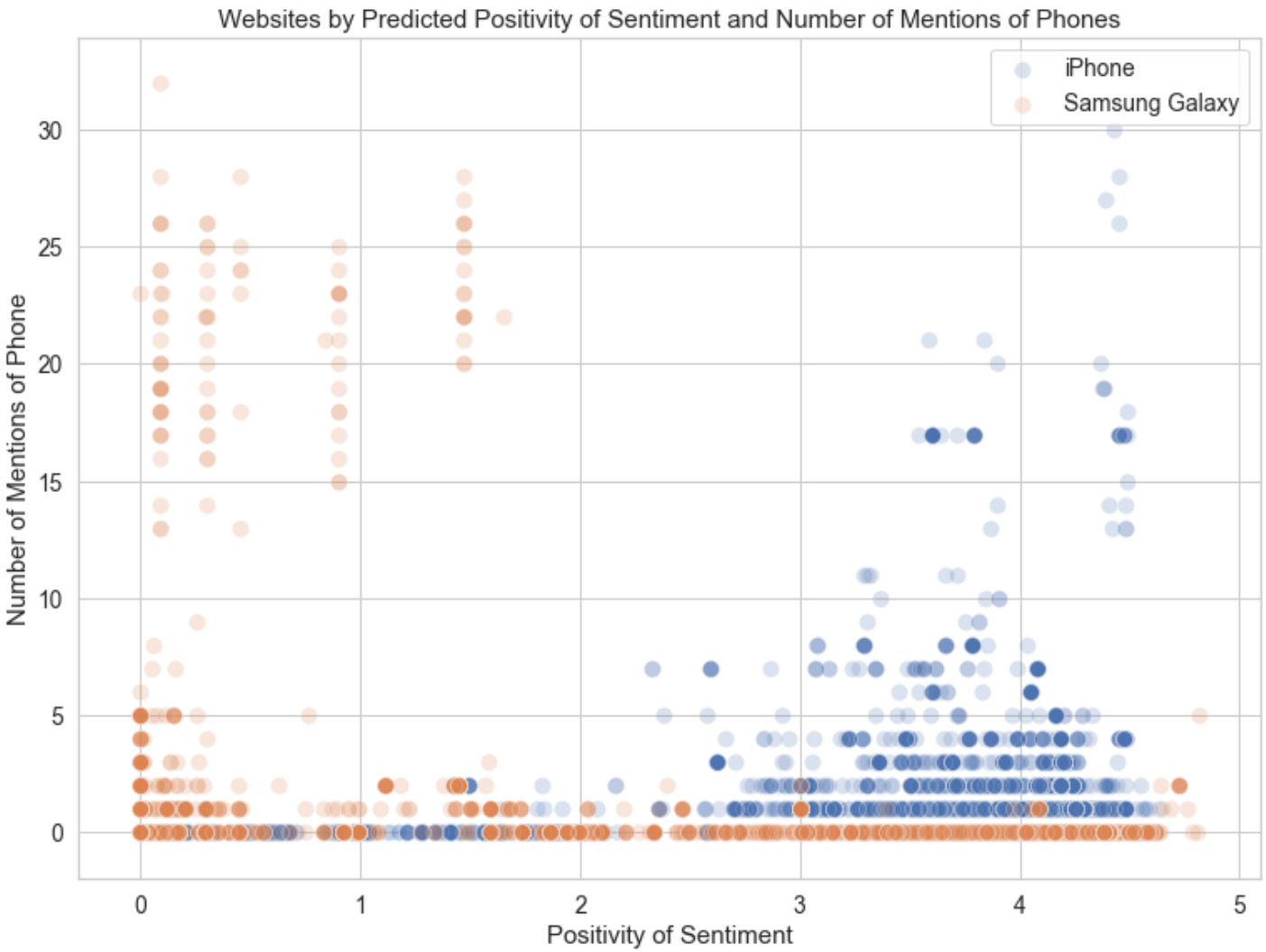
What To Do Better Next Time?
- Better Training Data
- Aggressive Limitations on What Sites to Crawl

Conclusions

Summary of Findings

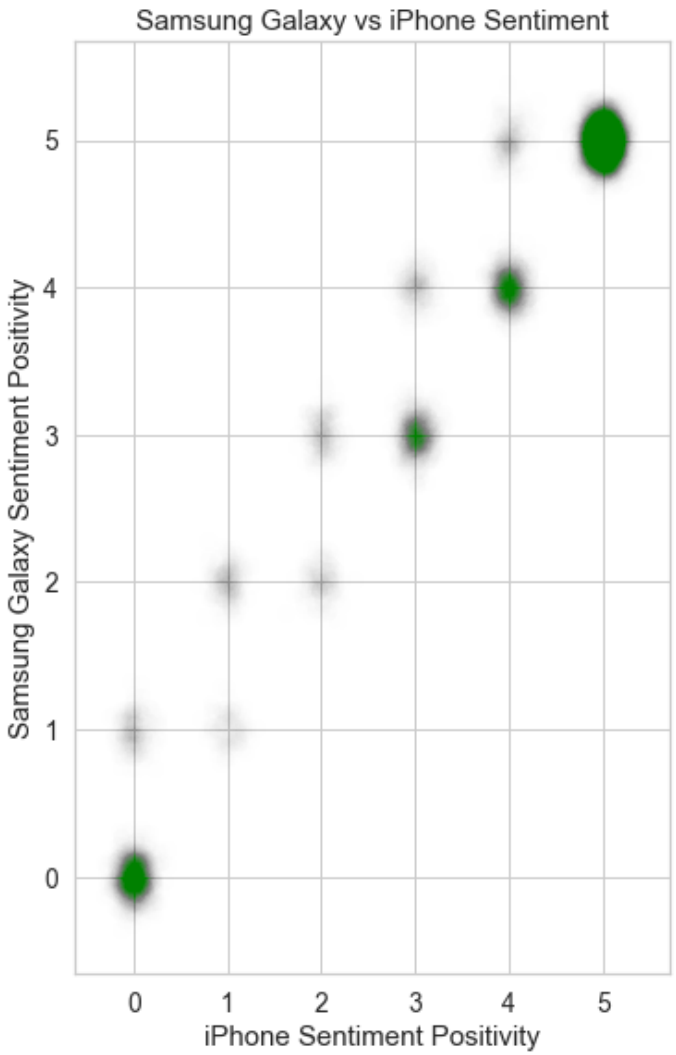
People Like iPhone More

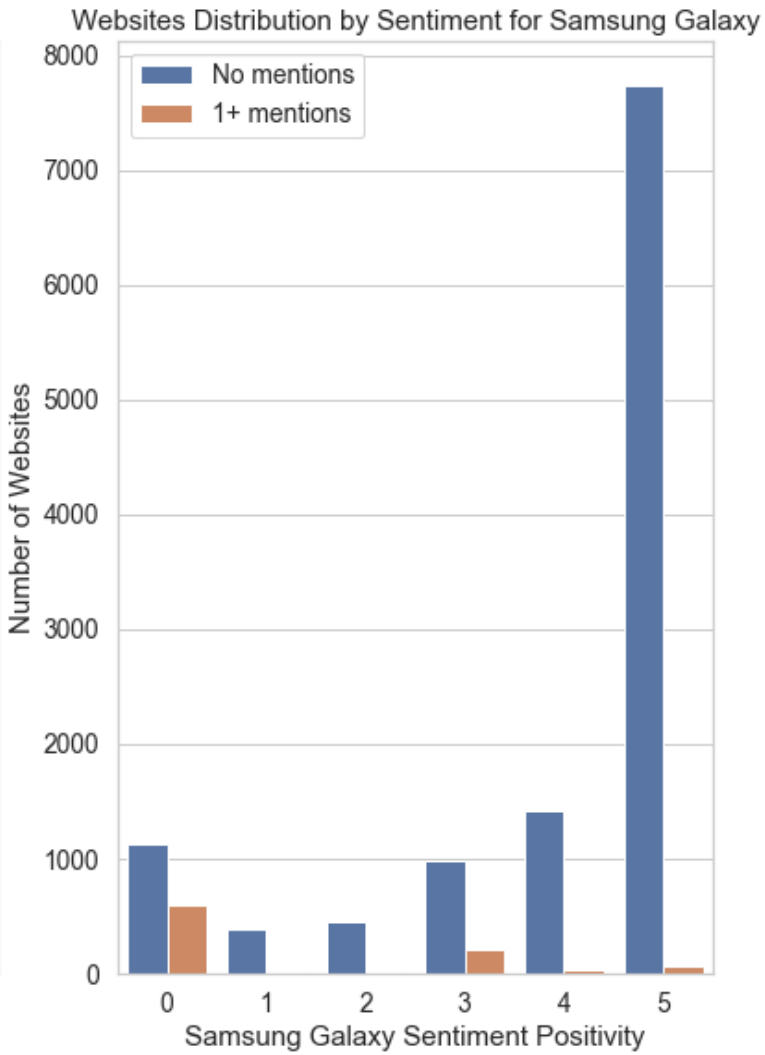
- When the sites actually mention the phones, iPhone sentiment is much more positive



. . . but the Training Data is Sketchy

- Sentiment labels for Samsung Galaxy might be derived from iPhone





. . . and the Websites Mentioning iPhone are Weird

- Danhostel is a danish hotel chain
- http://brazoriacountytexas.tk/ is hyperlink maze without any real content

What We Did?

Count Instances of Word and Word Combinations from Websites

1. Create scripts that count instances of words related to the the two phones in text files

Count Instances of Word and Word Combinations from Websites

1. Create scripts that count instances of words related to the the two phones in text files
2. Apply the scripts to a large number of websites taken from Common Crawl (cloud computing very useful here)

Count Instances of Word and Word Combinations from Websites

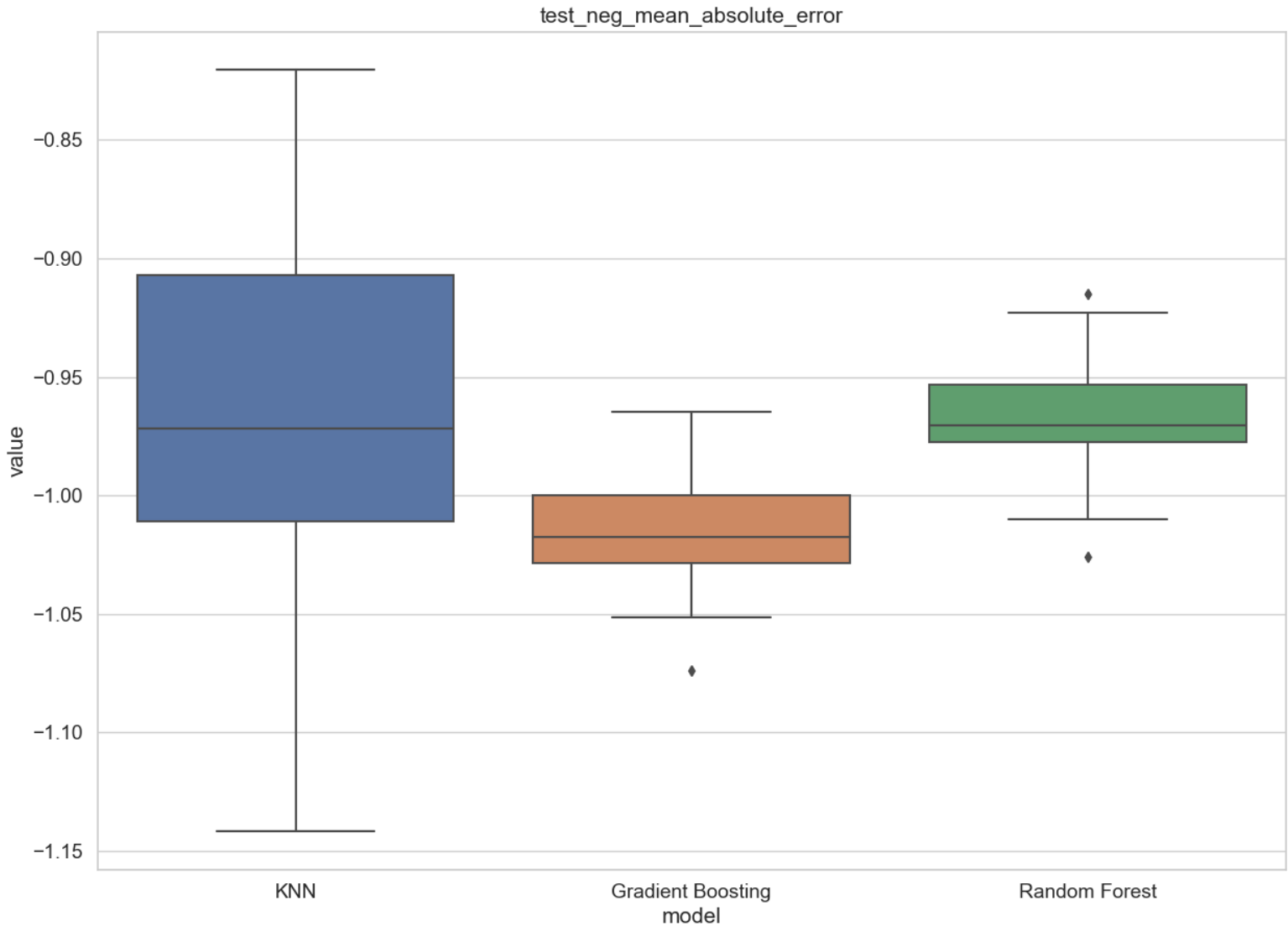
1. Create scripts that count instances of words related to the the two phones in text files
2. Apply the scripts to a large number of websites taken from Common Crawl (cloud computing very useful here)
3. Manually label part of the dataset for the sentiment towards the phones

Count Instances of Word and Word Combinations from Websites

1. Create scripts that count instances of words related to the the two phones in text files
2. Apply the scripts to a large number of websites taken from Common Crawl (cloud computing very useful here)
3. Manually label part of the dataset for the sentiment towards the phones
4. Predict the manually labeled sentiments from the number of word instances

Modelling

- Random Forest model perform accurately and is very stabile



What To Do Better Next Time?

Better Training Data

1. Better Documentation
   - Was the labeling really done by hand and if so, what were the guidelines for evaluation?

Better Training Data

1. Better Documentation
   - Was the labeling really done by hand and if so, what were the guidelines for evaluation?
   - If done programmatically, where is the code?

Better Training Data

1. Better Documentation
   - Was the labeling really done by hand and if so, what were the guidelines for evaluation?
   - If done programmatically, where is the code?
2. Check for Errors
   - Why are the sentiment almost identical for the two phones?

Better Training Data

1. Better Documentation
   - Was the labeling really done by hand and if so, what were the guidelines for evaluation?
   - If done programmatically, where is the code?
2. Check for Errors
   - Why are the sentiment almost identical for the two phones?
   - How can sites that don't mention phones have sentiments about them?

Aggressive Limitations on What Sites to Crawl

1. Think first what kind of sites do we want to include
   - Reviews and news probably easy to find

Aggressive Limitations on What Sites to Crawl

1. Think first what kind of sites do we want to include
   - Reviews and news probably easy to find
   - Forums are hard to interpret

Aggressive Limitations on What Sites to Crawl

1. Think first what kind of sites do we want to include
   - Reviews and news probably easy to find
   - Forums are hard to interpret
   - Beware of clickbait and ads

Aggressive Limitations on What Sites to Crawl

1. Think first what kind of sites do we want to include
   - Reviews and news probably easy to find
   - Forums are hard to interpret
   - Beware of clickbait and ads
2. Test that crawling really finds these sites
   - Wordclouds of the url

Aggressive Limitations on What Sites to Crawl

1. Think first what kind of sites do we want to include
   - Reviews and news probably easy to find
   - Forums are hard to interpret
   - Beware of clickbait and ads
2. Test that crawling really finds these sites
   - Wordclouds of the url
   - Manually visiting the sites

Conclusions

1. Don't trust the current findings. Make the decision by other means

Conclusions

1. Don't trust the current findings. Make the decision by other means
   - Sentiment for iPhone in sites actually mentioning the phone is quite positive, but this is hard to compare when we can't really trust the results for Samsung Galaxy

Conclusions

1. Don't trust the current findings. Make the decision by other means
   - Sentiment for iPhone in sites actually mentioning the phone is quite positive, but this is hard to compare when we can't really trust the results for Samsung Galaxy
2. More carefully documented labeling and data creation process in the future

Conclusions

1. Don't trust the current findings. Make the decision by other means
   - Sentiment for iPhone in sites actually mentioning the phone is quite positive, but this is hard to compare when we can't really trust the results for Samsung Galaxy
2. More carefully documented labeling and data creation process in the future
3. Web crawling should be more targeted

The End

Questions?