SentiLens

Tina Vu
20231216

Uncover review’s hidden sentiment

Understand thousands reviews at a glance
Know what reviews are about
More informative than simple star ratings

Impact

Quickly find
the preferred products
Fewer unexpected surprises

More satisfied consumers

Reduce wastage

Reduction in returns
More effective sales
Fewer out-dated stocks

Understand competitor's
advantage
Identify strength & weaknesses
in demand & supply

Better informed providers

Higher profitability

Higher conversion rate
Larger basket size

Dataset

Single sentence review

- Aspect term
- From & to char index
- Polarity

TEXT COLUMN:

I charge it at night and skip taking the cord with me because of the good battery life.

ASPECTS COLUMN:

[{'term': 'cord', 'polarity': 'neutral', 'from': '41', 'to': '45'},

{'term': 'battery life', 'polarity': 'positive', 'from': '74', 'to': '86'}]

Preprocessing

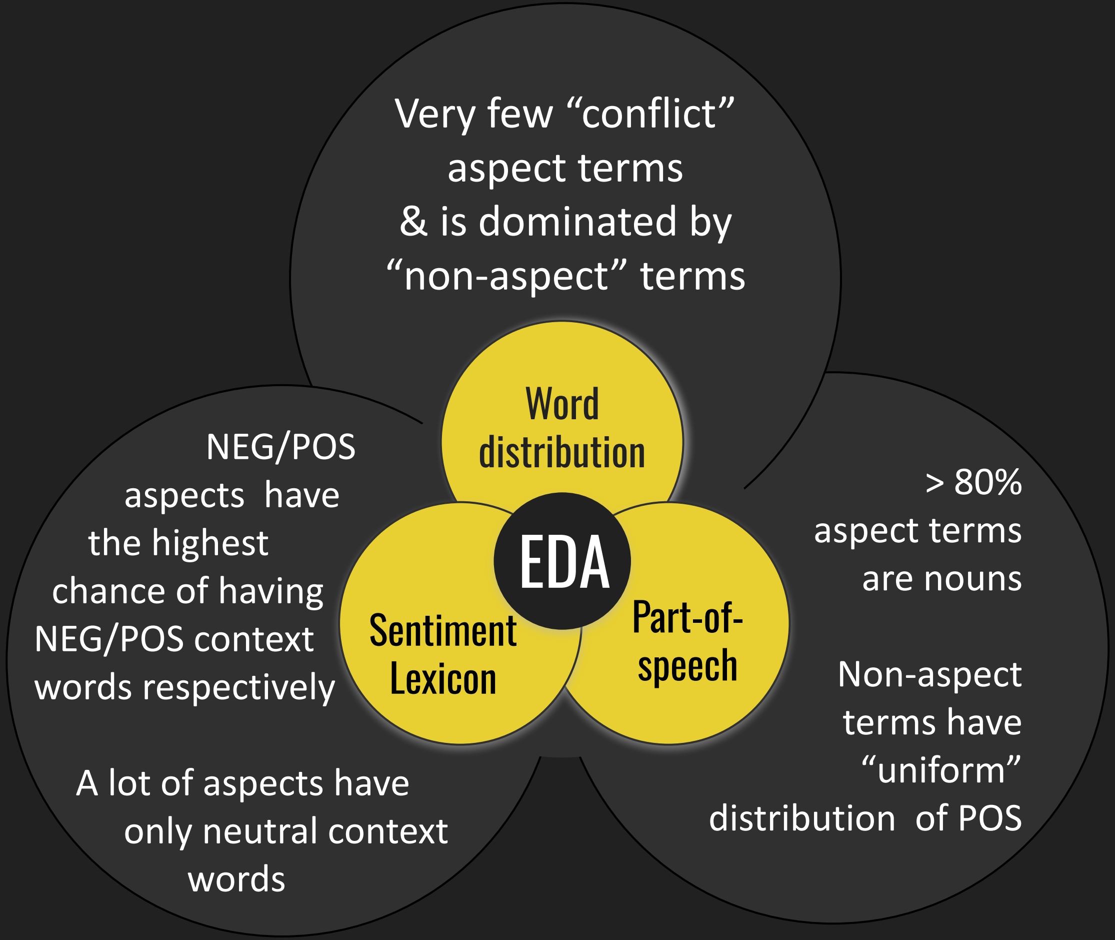
Unified-BIO tagging

Define target label using unified-BIO technique: aspect word boundaries and sentiment.
E.g. “Laptop <B-POS> screen <I-POS> is <O> amazing <O> . <O>”

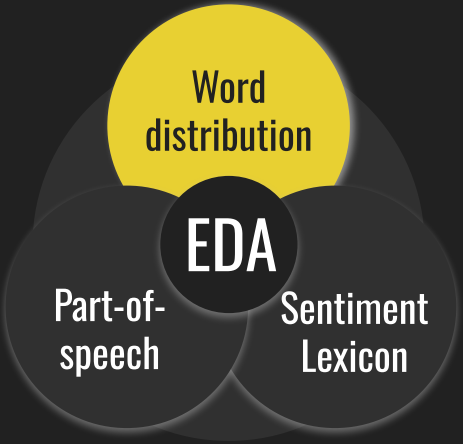
Build word features

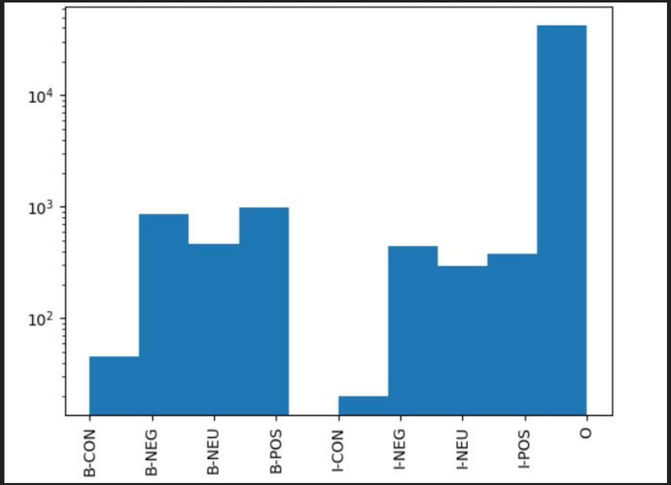
Incorporate information about word’s and its context words’ properties:

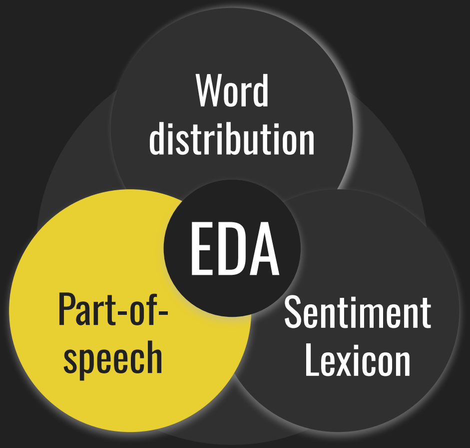
- Word index & reverse index in the sentence
- Part-of-speech
- Sentiment lexicon, …


Very few "conflict" aspect terms & is dominated by "non-aspect" terms
Word distribution
NEG/POS aspects have the highest chance of having NEG/POS context words respectively
A lot of aspects have only neutral context words
Sentiment Lexicon
EDA
Part-of-speech
> 80% aspect terms are nouns
Non-aspect terms have "uniform" distribution of POS

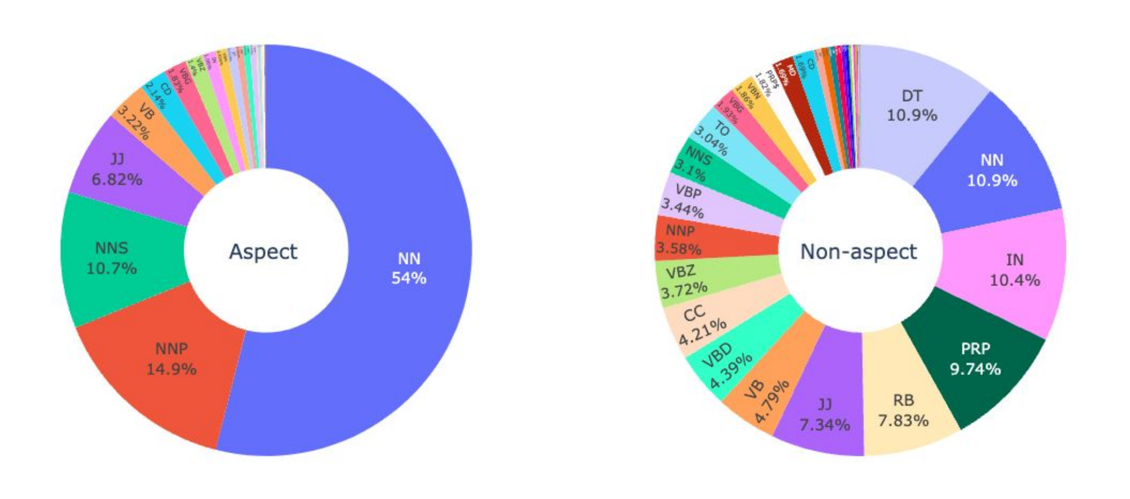
Very few “conflict” aspect terms
& is dominated by “non-aspect” terms



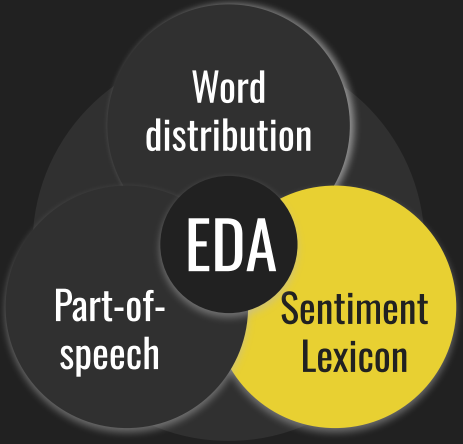




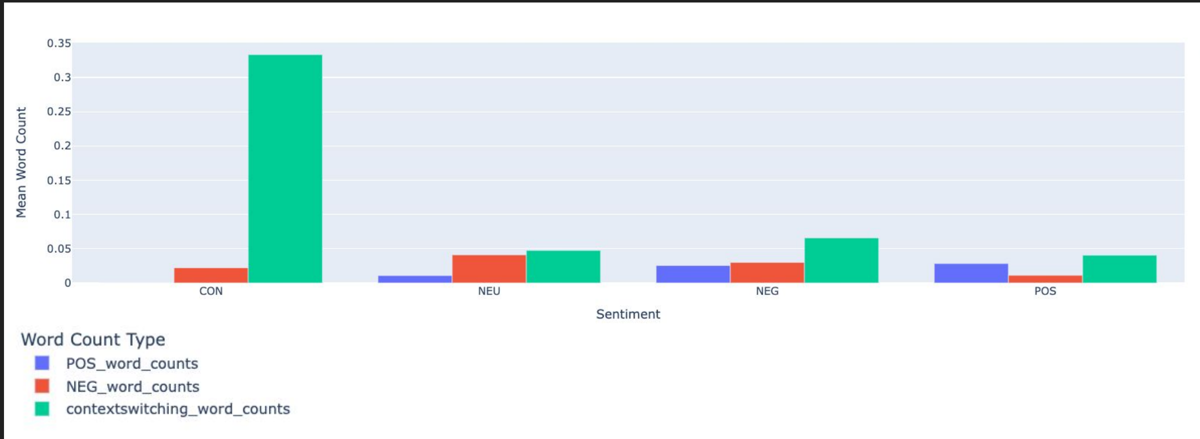
> 80% aspect terms are nouns
Non-aspect terms have “uniform” distribution of POS



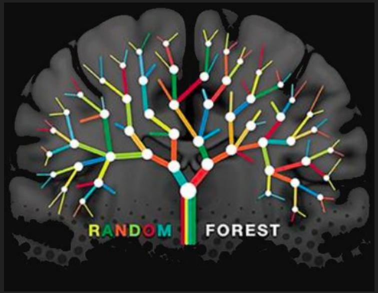
* POS starts with N are nouns



- A lot of aspect terms do not have pos/neg context words
- >35% “conflict” terms have context switching context words
- Whichever term polarity, they may contains both “positive” & “negative” context words, just the ratio between different sentiment lexicons are slightly different.



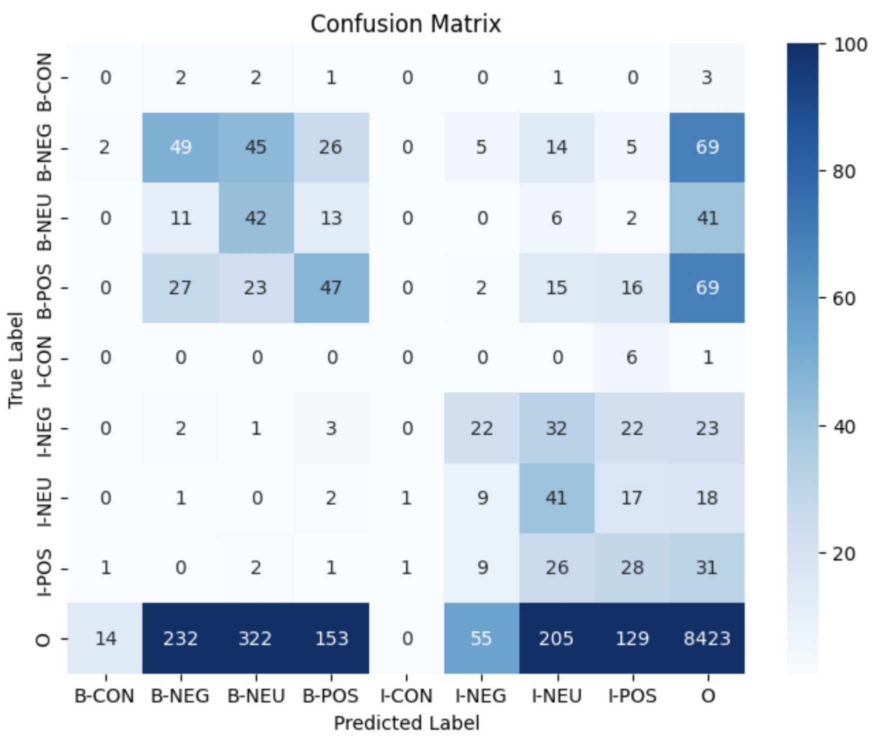
Random forest



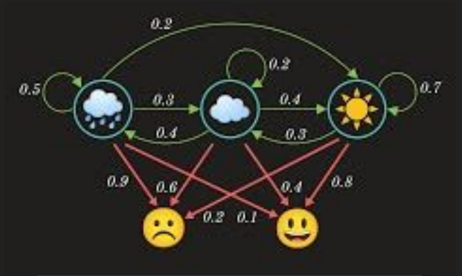
Macro avg f1-score: 0.22

A lot of words were misclassified as “O” (non-aspect term)

There seems to be mix-ups in polarity, while not much in “B” & “I”



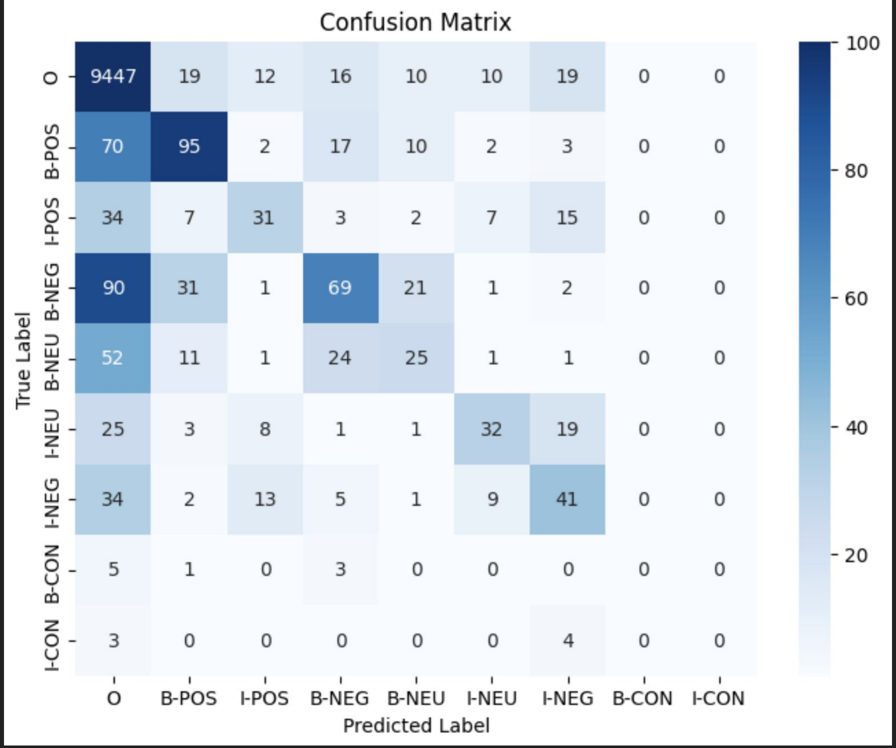
Conditional Random Field



Macro avg f1-score: 0.37

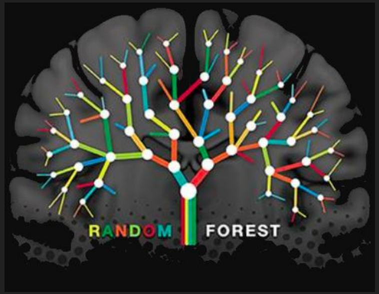
A lot of words were misclassified as “O” (non-aspect term)

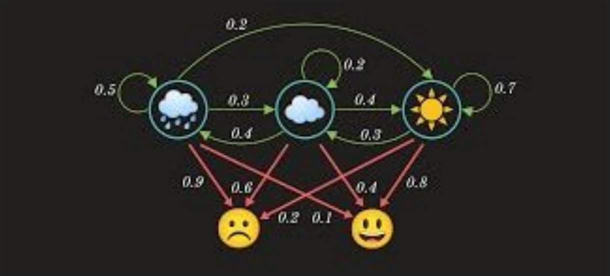
The model could not predict any “conflict” aspects, and is less likely to predict aspect as non-aspect.



Baseline model

Macro average f1-score





Random forest:
0.22

Conditional Random Field:
0.37

Next step

Feature improvement:

Add more features: head words, Google Word2Vec word embeddings

Unsupervised training

Re-train model using rule-based aspect term extraction on larger dataset

Productionize:

An app to allow users to look for specific product reviews & filter review by aspects + summarize aspect sentiments