1 intro: 150

1. Describe the raw dataset
2. Describe the general process

2. preprocessing:

a. data cleaning: Aim: describe the benefit, comparison between the data before cleaning and after cleaning

lots of noise, could reduce the potential dimension

irrelevant to sentiment: special words (@username, #tag, URL), special characters, punctuation, numbers, stop words,

the some word with different forms: contraction, accented\_chars

different words have the some lexeme

b. vectorization(transformation)

BOW vs TFIDF

choose of ngram

1. Comparison

Reduce dimension: 204657 -> 127482

3. feature selection

To further reduce dimension

Chi2 vs f\_classifer vs MI

comparison

4. training:

a. Split data for validation

reason

random holdout vs k fold

b. modeling

Base model: 0R

~~delete words are irrelevant with sentiment~~

~~accented~~\_chars: reduce dupulicate: the word with or without accented chars are mostly the same

~~contraction~~: like I’m -> I am, because the full text and the contraction have the same meaning, should be classified into one attributes

~~convert\_num: irrelevant with sentiment~~

extra\_whitespace=True,

lemmatization=True,

~~lowercase=lowercase words and uppercase words have the same meaning~~

~~punctuations=irrelevant with sentiment~~

~~remove\_html=True,~~

~~remove\_num=True,~~

~~special\_chars=True,~~

~~stop\_words=irrelevant to sentiment, high frequency~~