Pre-Processing issues:

- Vocab size: ~26K from just ~17K entries

  Learning custom embeddings for the vocabulary of this magnitude provides more scope to overfit.

- Lack of punctuations (← default tokenizers)

  Punctuations were understood to be important in certain cases of positive classes. (is definition) based on our manual analysis over the data

- More ambiguity in samples (Example Sentences)

  The guidlines for definition were not clear in certain cases. Some examples from Notes.md

Feature solution

- Meaningful replacement for OOV

  Replace with pos Tags instead of <unk>

- Pos with Punctuations

  Using "PUNCT" instead of all punctuation is ambiguious. So punctuations from the set indicated in Notes.md are used as is.

- Dependency feature:
  Concatenation of modifier, Head and label representations.

- Tokens + Pos w punct
  Concatenation of Token rep + pos rep [Our Inclusion updated to these Paper work]

Assumption:

Syntactic is all we need:

Initial assumption: Semantics might not have any effect for this Classification task
Turns out: Using Pretrained word vector has a positive impact

Todo: May need to run an Exclusive Exp to support this case.