TOKENIZER :

Input : An English paragraph consisting of Words, Emoticons and other punctuation( comma, full stop ) ?Usernames

Output : An array of arrays which in turn consist of words

Steps :

Sentence tokenize

Word tokenize (White space )

Then separate full stand-alone punctuations,