Enhancing Emotion Recognition Using POS Tagging

Assignment - 1

NLP CS60075
Autumn Semester 2024
IIT Kharagpur

TASKS

01 POS Tagger Implementation

[40% of TOTAL]

02 Vanilla Emotion Recognizer

[15% of TOTAL]

03 Improved Emotion Recognizer

[25% of TOTAL]

04 Report

[20% of TOTAL]

Sample Code -
https://colab.research.google.com/drive/1sfARw_asCMJwiFIGIUBUernwnuC8uoaN?usp=sharing

POS Tagging

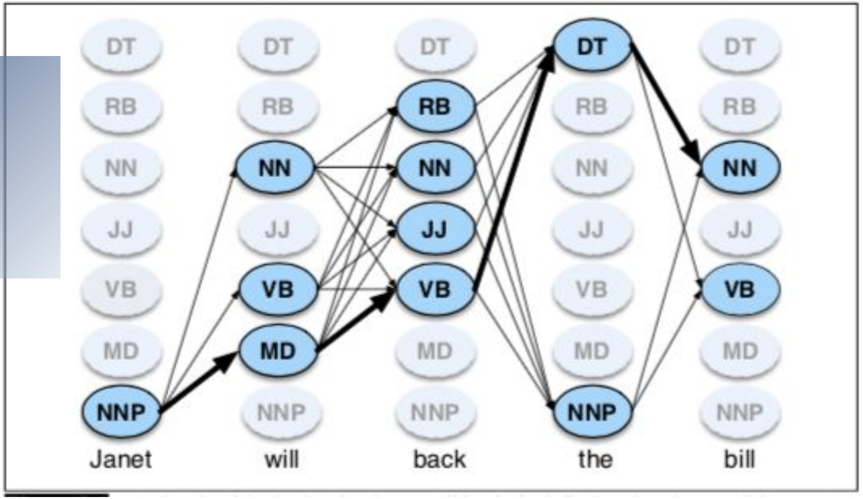
Natural Language Tagging or more specifically Part-of-Speech (POS) tagging, involves assigning grammatical categories (such as nouns, verbs, and adjectives) to words in a sentence.

The impact of this primary linguistic task extends far beyond syntax:

- Accurate POS tagging contributes to understanding the context, semantics, and even emotions conveyed by the text.
- We have seen that some words can be used as multiple parts of speech in the English language. For example:

“Kevin has dark hair and fair skin.” (ADJECTIVE)

“The new fair is boring.” (NOUN)



VITERBI ALGORITHM

The Viterbi algorithm is a dynamic programming algorithm for obtaining the maximum a posteriori probability estimate of the most likely sequence of hidden states—called the Viterbi path—that results in a sequence of observed events, especially in the context of Markov information sources and hidden Markov models (HMM).

TASK 1:

POS Tagger Implementation (from scratch)

DATA

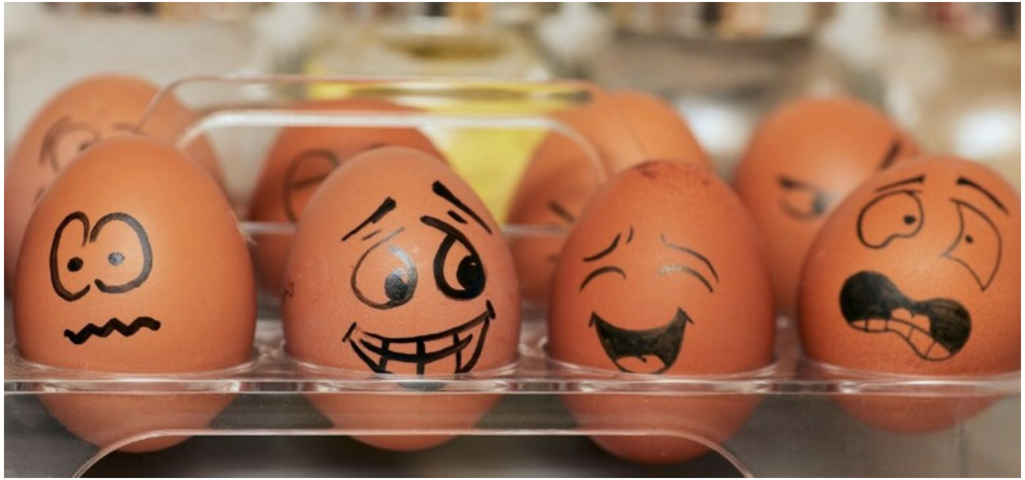
Use the treebank corpus (from nltk) for training data. [nltk downloader]

IMPLEMENT

Implement the Viterbi Algorithm (dynamic programming) for POS Tagging

POINTERS

You can keep any number of POS Tags, but make sure your set contains NOUN, ADJECTIVE, VERB.



WHAT IS EMOTION RECOGNITION?

The process of computationally identifying and categorizing emotions expressed in a piece of text, especially in order to determine the writer's state of mind towards a particular topic, product, etc. Some typical emotions incorporate joy, sadness, anger, fear, love and surprise.

Vectorization

Vectorization in NLP refers to the process of converting textual data, which is inherently unstructured and composed of words, sentences, or documents, into numerical vectors. These numerical representations enable machine learning algorithms to work with text data,

Example:

- ❑ **TFIDF** [Use this vectorizer for the assignment]
- ❑ Word2Vec
- ❑ Bert Embeddings

$$w_{x,y} = tf_{x,y} \times \log\left(\frac{N}{df_x}\right)$$

TF-IDF

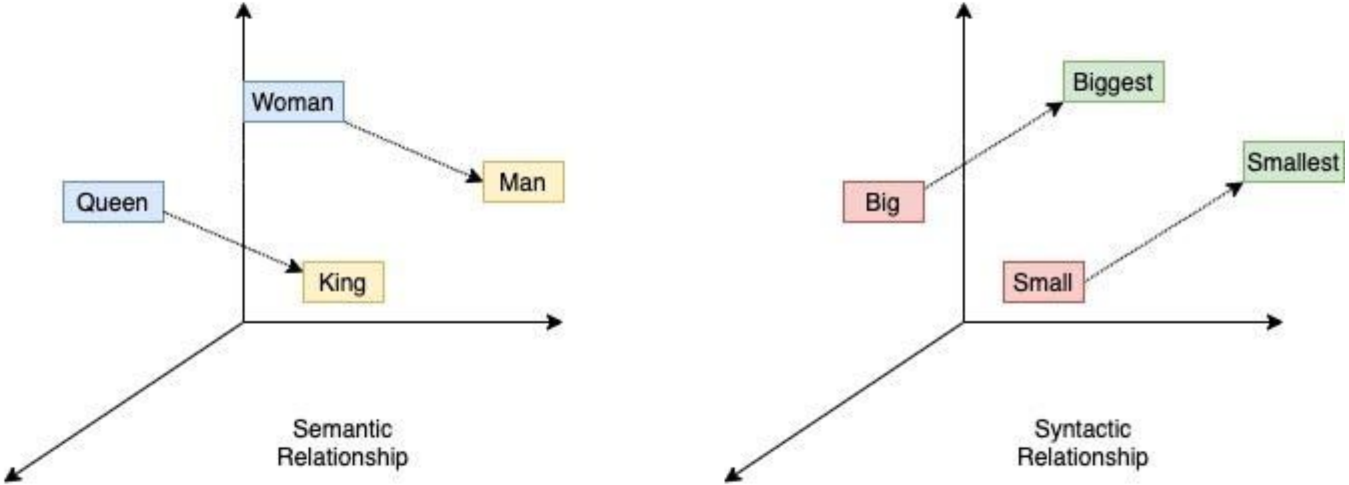
Term x within document y

$tf_{x,y}$ = frequency of x in y

df_x = number of documents containing x

N = total number of documents

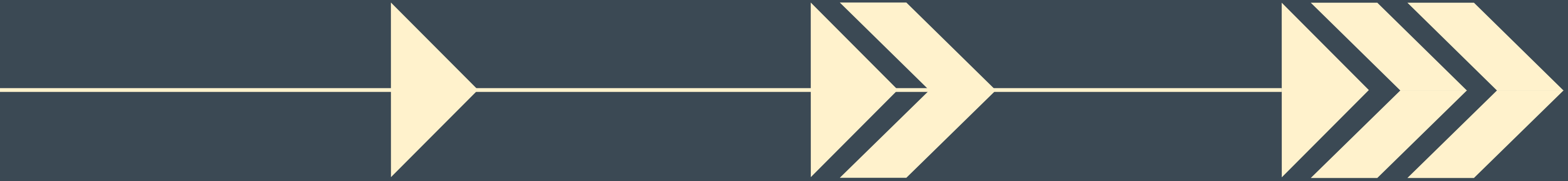
Word2Vec



Capturing Semantic as well as Syntactic Relationship between words

TASK 2:

Vanilla Emotion Recognizer



DATA

Use the twitter_messages corpus for data. [Use data from above provided link only] [See instructions for more details]

VECTORIZER

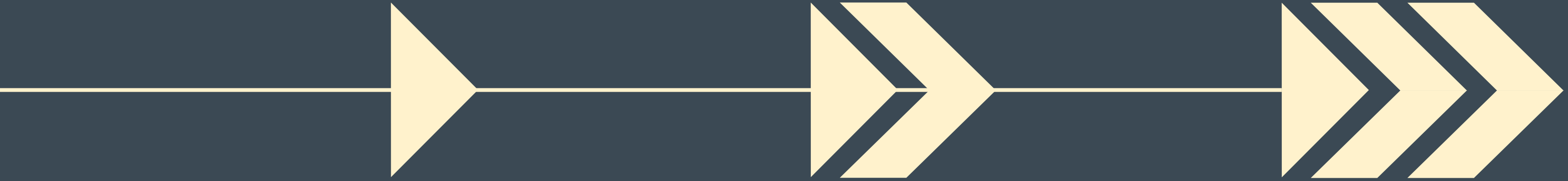
You have to use Tf-idf vectorizer

TRAIN

Train any Classifier (Naive Bayes, SVM, etc.) for emotion recognition using the above features

TASK 3:

Improved Emotion Recognizer



INTEGRATE

Use the POS Tagger in Task 1 for POS tagging the dataset.

PIPELINE

Implement a pipeline to integrate the POS tag features along with the sentence embeddings. (Be creative)

TRAIN

Train the same Classifier (as chosen in Task 2) again for emotion recognition using the new features

REPORT AND SUBMISSION GUIDELINES

REPORT

- ❑ Add your observations to a report (submit in pdf format).
- ❑ Compare the performance of your POS-tag-enhanced model with a baseline model that doesn't use POS tags. [Don't worry about scores]
- ❑ Make sure to include the classification reports and confusion_matrix on test split of both models in the report.
- ❑ Make sure to highlight any advanced modifications that you've done in your report

SUBMISSION GUIDELINES

- ❑ You have to use IPython Notebooks for coding. (Use Google Colab, Kaggle for running your assignments).
- ❑ Make sure your IPython Notebook has the outputs from each cell. The classification report and confusion matrix must be present in the submitted notebook as well. Absence of these results will lead to deduction in marks.
- ❑ Refer to instructions for more details.

Assignment Deadline: 11:59 PM, 3rd September 2024

HAPPY LEARNING!

Naquee Rizwan
CSE, PhD Scholar