Assignment Discussion #3 and Project Discussion

Singamsetty Sandeep, 213050064
Anmol Namdev, 213050044
Sakharam Sahadeo Gawade, 213050027

12th October, 2021

Assignment Discussion #3

WSD continuation, NEI

Performance report of HMM-WSD

On 5 folds Cross-Validation

Earlier with HMM: 0.501413 ± 0.000475

Improved preprocessing for HMM and added MFS and WFS for unseen words

HMM+WFS:0.576162 ± 0.001714

HMM+MFS:0.578440 ± 0.001010

Problem Definition: NEI

- Given a sentence/document, mark each token as 1/0 as per whether the token is a Named Entity or not
- If the named entity consists of multiple words just continue with 1s until a non-NE appears
- E.g. *The_0 State_1 Bank_1 of_1 India_1 is_0 the_0 largest_0 bank_0 in_0 the_0 country_0 . _0*

DATA

- CoNLL 2003
- Dataset obtained from huggingface and the conversion of named entity tags was already converted to numerical form.
- Tokens' NER tag was in [0,8], converted to binary such that 0 is not a named entity and 1 is named entity

Feature Engineering

- f1: First character is capital?
- f2: Is a stop word?
- f3: Length of the word
- f4: No. of capital letters
- f5: No. of dots
- f6: No. of hyphens
- f7: No. of numbers
- f8: No of Letters
- f9: POS tag (in numerical form)
- f10: Is proper noun?
- f11: Is a verb?
- f12: Is a cardinal number?
- f13: Previous is NE?
- f14: Previous article?
- f15: Is the first word?

Justification of Feature Set

- Knowing if first character is capitalized helps in capturing ENAMEX
- POS tag features used to improve identification of ENAMEX, NUMEX and TIMEX
- To increase the number of vectors, we inculcated integer valued features such as the length of the word, no of symbols,...
- Some sentences were completely in capital letters so to distinguish them from the words with only first character in capital, length of word and number of capital letters can together help
- No. of hyphens, numbers and letters along with the length of the word can help in identifying TIMEX and NUMEX
- Detecting stop-words will help to identify the non named entities, also can be useful in features that incorporate context

Performance

- Overall Precision: 0.9530
- Overall Recall: 0.9530
- Overall F-score: 0.9530
- Overall Accuracy: 0.9530

	precision	recall	f1-score	support
0	0.9926	0.9501	0.9709	38323
1	0.8041	0.9667	0.8779	8112
accuracy			0.9530	46435
macro avg	0.8983	0.9584	0.9244	46435
weighted avg	0.9597	0.9530	0.9547	46435

Confusion Matrix

Predicted □ Actual (rows)	0	1
0	36412	1911
1	270	7842

RBF Kernel Used

```
svm.SVC(cache_size=10000,  class_weight="balanced",
                C=0.1,gamma=10)
```

Weight balancing attribute to obtain the result, similar result when using SMOTE

Similar results with other values of C and gamma

Result Interpretation

- There are a lot of a false positives(non named entity classified as named entity) even after balancing.
- Sklearn's Linear SVM returns class weights which can be used to used know the most important features.

```
svm.LinearSVC(penalty='l1',
dual=False,class_weight="balanced",C=0.1)
```

f1: First Letter Capital

f8: No. of letters and

f10: Is proper noun? had higher weights compared to others

- With higher computing resources and long resource time, more features capturing context can be included

Project

Tweet Classification for Crisis Response

What is the “Problem”

- Classification of Crisis Tweets
- Comparing models across different disasters and an aggregated dataset
  - Input: Tweets
  - Output(Multi-class classification):
    i. Critical Disaster Tweets
    ii. Non Critical Disaster Tweets
    iii. Non Disaster Tweets

Why is the problem important

- Twitter is medium where information is shared in a concise manner.
- Critical Information and Insight can be gained from the tweets which can be used to make a speedy rescue plan
- For this it is not just important to know if the tweets are related to a disaster or not but also to distinguish between critical and non critical tweets.

What is hard about the problem

- Tweets may have unclear context due to the word limit.
- Preprocessing of Tweets is challenging because of the use of acronyms(Out Of Vocabulary Words), spelling mistakes, use of english alphabet to write in non english sentence, use of sarcasm
- Dealing with false positives
- Might have high computational needs for training
- A model trained on one disaster dataset may perform better than on an aggregated dataset *

What has been done on this problem so far

1) Identifying and Categorizing Disaster-Related Tweets by Stowe et. al. ,"Proceedings of The Fourth International Workshop on Natural Language Processing for Social Media",2016. ACL
2) A Sentiment-Aware Contextual Model for Real-Time Disaster Prediction Using Twitter Data by Song et. al., Future Internet, Volume 13, 2021
3) Deep Learning and Word Embeddings for Tweet Classification for Crisis Response by ALRashdi et al.,The 3rd National Computing Colleges Conference, 2018
4) Cross-Lingual Disaster-related Multi-label Tweet Classification with Manifold Mixup by Ray Chowdhury et. al., Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, 2020. ACL
5) Natural Language Processing to the Rescue?: Extracting “Situational Awareness” Tweets During Mass Emergency, Verma et. al. Proceedings of the Fifth International Conference on Weblogs and Social Media, 2011. AAAI

Your tackling of the problem

- Procedure to solve the problem:
  - Preprocessing
  - Tokenization
  - Word Embedding Generation
  - Model Building and Training
  - Comparison of Models across disasters
- Our project is based on [1], [2] & [5]
- Resources you will need: CrisisNLP data, huggingface(XLNet), Keras(Bi-LSTM), scikit learn(SVM) Computing facility: _____
- Performance Metrics: Precision, Recall and F-Score

Models for Performance Comparison

EARTHQUAKE	FLOODS	PANDEMIC	HURRICANE	AGGREGATED

Crisis_word2Vec + SVM
Crisis_word2Vec + BiLSTM
Crisis_word2Vec + BiLSTM + CNN
SentiBERT + SVM
SentiBERT + BiLSTM
SentiBERT + BiLSTM + CNN
XLNet + SVM
XLNet + BiLSTM
XLNet + BiLSTM + CNN

Thank You