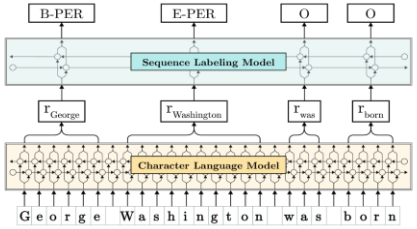


GEOAI CHALLENGE

Location Detection from Social Media Crisis-related Text

MASTER OF SCIENCE IN
BUSINESS ANALYTICS AND DATA SCIENCE
Spears School of Business

MASTER OF SCIENCE IN
BUSINESS ANALYTICS AND DATA SCIENCE
Spears School of Business

Patterns

Saswata Rautray

Prathamesh Kulkarni

Noreen Chihora

Tejaswi Maruthi

Becka Cammon

MASTER OF SCIENCE IN
BUSINESS ANALYTICS
AND DATA SCIENCE
Spears School of Business

Patterns

About ourselves

We are team patterns, a team of ML enthusiastic second-year student of MS Business Analytics & Data Science at Oklahoma State University.

We have keen interest in NLP and text analytics.

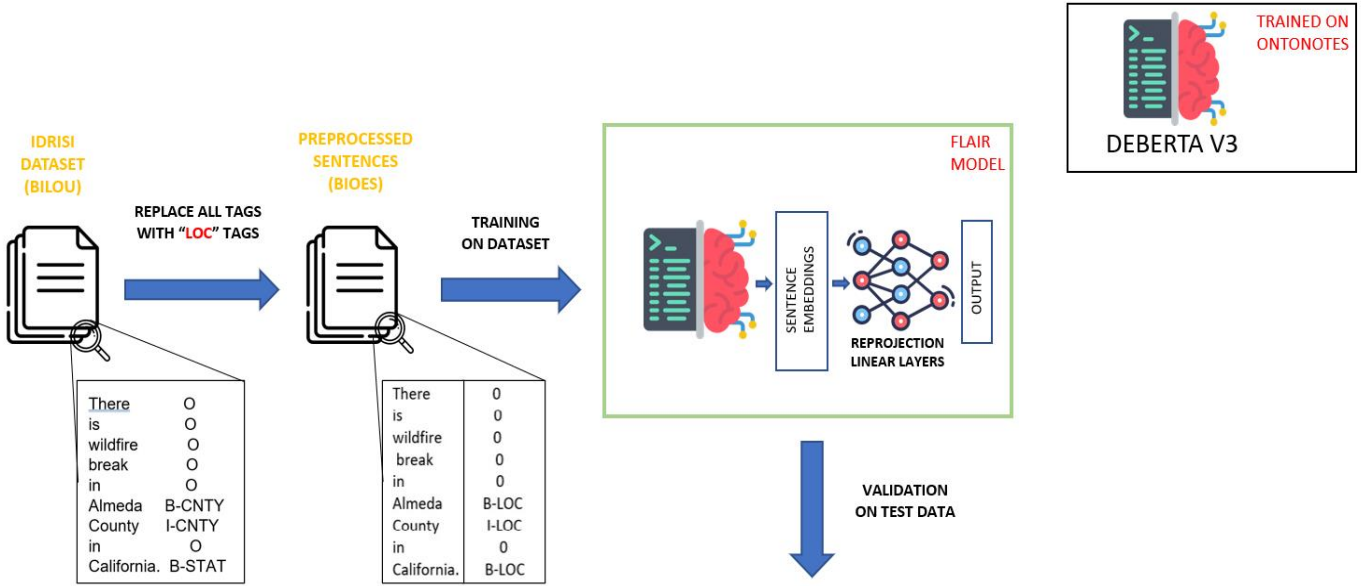
Mission Statement

Using Machine Learning to solve real world problems and make earth a better place to live.

Agenda

- Introduction
- Methodology
- Evaluation
- Takeaway messages

Methodology




```
Sentence: "George Washington went to Washington ." → ["Washington"/LOC]
The following NER tags are found:
Span[4:5]: "Washington" → LOC (0.9703)
```

Methodology

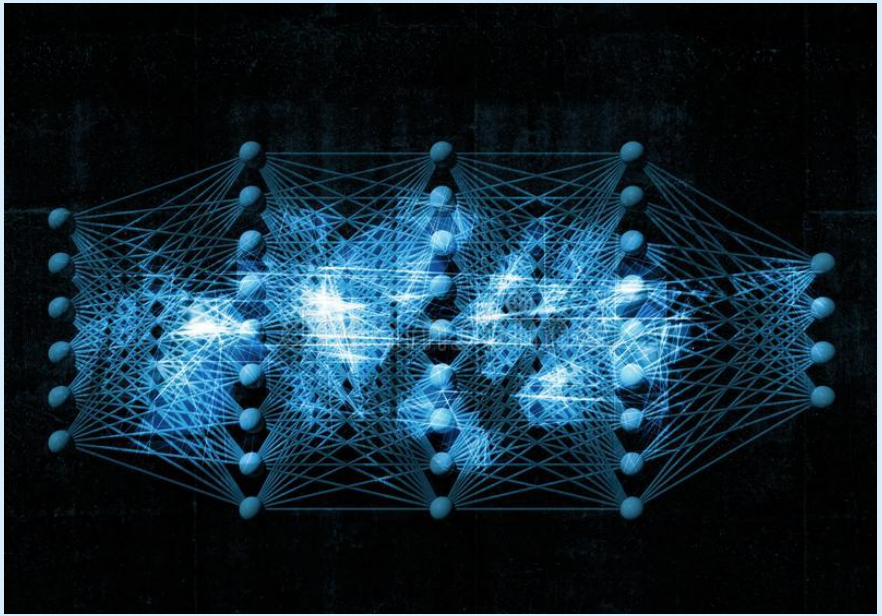
Hypothesis and assumption:-

Our main hypothesis was that transfer learning can better help to tackle challenge of NER task in NLP.

Our assumption was finetuning a pretrained embedding from a bigger corpus would better perform in understanding the context of given sentence.

Type of Method:-

ML based- Flair framework (HuggingFace)



Methodology

Evaluation:-
Programming language:- Python, Colab (GPU-resource)

Preprocessing steps:-
- ❑ Importing data from IDRISI in GitHub.
- ❑ Changing the tagging from “BILOU” to “BIOES”.
- ❑ Changing multiple tags like STAT, CNTY to “LOC”.

LMR model:-
- ❑ Using the Flair framework from “HuggingFace”.
- ❑ Deberta V3 pretrained on “OntoNote5”.
- ❑ Adding a linear layer for reprojecting the embeddings.
- ❑ Adding a linear layer for feature detection
- ❑ Optimizing the loss function by using weights.

Off-the-shelf adopted tools:-
- ❑ Using the easy to use NLP framework from HuggingFace called “Flair”.

Evaluation

Training & Development:-

- The pretrained embeddings were used from deberta model that was trained on “OntoNote5”.
- Input data Schema:- The input data was from IDRISI-R GitHub
- The data format was changed from “BILOU” to “BIOES”.
- Only single label “LOC” used for all location.
- Total training data 14392 and validation data is 2056.

Model Tuning :-

- Hyper parameters:-
- learning rate:- 5.0e-6
- Reproject embedding:- True
- Mini batch size:- 7
- Epochs:- 3
- Optimizer:- AdamW
- Weight decay: - 0.
- Hidden state:- 768

Evaluation

As we can see that Flair based DeBerta model performs well for the disaster-prone dataset.

Flair-DeBerta based model has a F-1 score of 0.902 while BERT and CRF model have score of 0.889 and 0.864.

Therefore, our system better performs than the existing LMR models.

Team (Run)	Flair -DeBerta			BERT-baseline Type less			CRF baseline Type less		
Event	P	R	F1	P	R	F1	P	R	F1
California Wildfires 2018	0.921	0.930	0.920	0.920	0.930	0.920	0.910	0.890	0.890
Canada Wildfires 2016	0.758	0.797	**0.765**	0.740	0.760	0.740	0.740	0.750	0.730
Cyclone Idai 2019	0.889	0.912	**0.893**	0.940	0.920	0.920	0.930	0.880	0.890
Ecuador Earthquake 2016	0.959	0.962	**0.957**	0.960	0.950	0.950	0.940	0.910	0.920
Greece Wildfires 2018	0.929	0.948	0.930	0.930	0.930	0.920	0.950	0.930	0.930
Hurricane Dorian 2019	0.886	0.912	**0.894**	0.870	0.890	0.870	0.860	0.850	0.850
Hurricane Florence 2018	0.816	0.811	**0.805**	0.800	0.780	0.780	0.770	0.730	0.740
Hurricane Harvey 2017	0.919	0.921	0.914	0.910	0.900	0.900	0.900	0.880	0.890
Hurricane Irma 2017	0.849	0.859	0.851	0.850	0.850	0.840	0.790	0.780	0.780
Hurricane Maria 2017	0.915	0.921	0.914	0.920	0.910	0.910	0.910	0.880	0.880
Hurricane Matthew 2016	0.959	0.950	**0.951**	0.960	0.940	0.940	0.940	0.890	0.900
Italy Earthquake Aug 2016	0.928	0.924	**0.925**	0.880	0.880	0.870	0.820	0.810	0.820
Kaikoura Earthquake 2016	0.920	0.930	0.920	0.920	0.920	0.910	0.880	0.870	0.870
Kerala Floods 2018	0.908	0.923	**0.907**	0.890	0.900	0.880	0.870	0.830	0.840
Maryland Floods 2018	0.889	0.942	0.906	0.920	0.900	0.900	0.930	0.890	0.900
Midwestern US Floods 2019	0.932	0.965	**0.938**	0.930	0.930	0.930	0.930	0.910	0.920
Pakistan Earthquake 2019	0.878	0.917	**0.885**	0.940	0.950	0.940	0.960	0.910	0.920
Puebla Mexico Earthquake 2017	0.932	0.949	**0.937**	0.890	0.910	0.890	0.900	0.890	0.880
Srilanka Floods 2017	0.934	0.940	**0.932**	0.900	0.900	0.890	0.900	0.850	0.870
Average	**0.901**	**0.916**	**0.902**	**0.898**	**0.897**	**0.889**	**0.886**	**0.859**	**0.864**

Takeaway Messages

Key Learnings:-

- ❑ Transfer learning can be used to a great extent effectively in NLP problems like entity recognition, Question answering, text summarization etc.
- ❑ Flair is great easy to use framework for NLP implementation.
- ❑ CRF related “Viterbi Loss” function can be outperformed using “CrossEntropy loss” if weights are provided for classes.

Future Scope:-

- ❑ Can use “ACE” or “automatic concatenation of embedding” for finding best suited embedding in flair. Combination of embeddings is generally said to give better result.
- ❑ Other hyperparameters can be explored more such as adding a RNN layer for feature input, subtoken_pooling, layers, fine_tuning etc.

Thank you !!