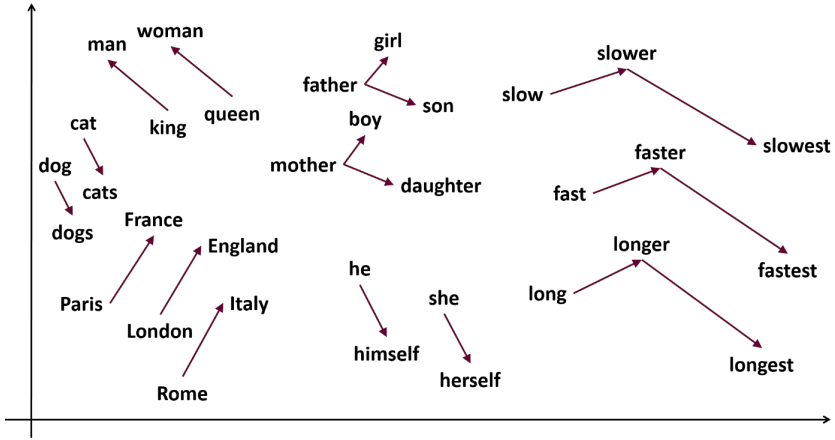
NLP Maps
Data Science Capstone Project 2023

Andrew Simon, Meenal Rawlani, Shijie Zhang

Introduction

- Embedding is a key step in Natural Language Processing (NLP), converting words to vectors in order to capture semantic relationships



Introduction

- Embedding is a key step in Natural Language Processing (NLP), converting words to vectors in order to capture semantic relationships
- NLPMaps leverages word embedding models to enhance safety in chemical plants by finding the best word embedding method to predict accident severity
- NLP Maps was able to determine the best embedding method for Dow accident reports out of 7 models and 4 classifiers
- Its modular algorithm is capable of being extended to any type of text classification

Data

01 **IMDB Dataset**

- Popular benchmark dataset for study of sentiment analysis
- 50k entries, 25k each for positive and negative reviews

02 **PSE Dataset**

- Chemical plant safety/incident reports from Dow
- Three kinds of reports: Same Person Report, Multiple People Report, and Multiple People Less Details. 100 entries each, total of 300
- Five levels of safety as labels

Models

Bag of Words

Represents text as a collection of individual words

Ignores word order and context

TF-IDF

Weighting scheme that highlights important terms in a document collection

Word2Vec

Efficient word embeddings that capture semantic relationships and analogies.

FastText

Enriches word embeddings by implementing character n-grams, improving performance for rare and out-of-vocabulary words

Models

GloVe

Word embeddings that capture global word co-occurrence statistics

Bert

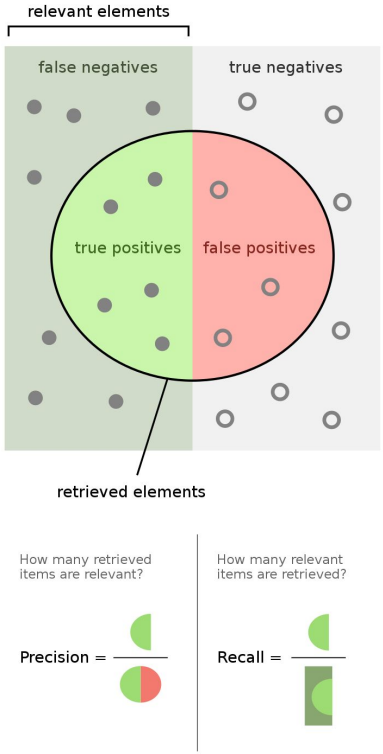
Pretrained transformer model with contextual deep bidirectional representations

ELMo

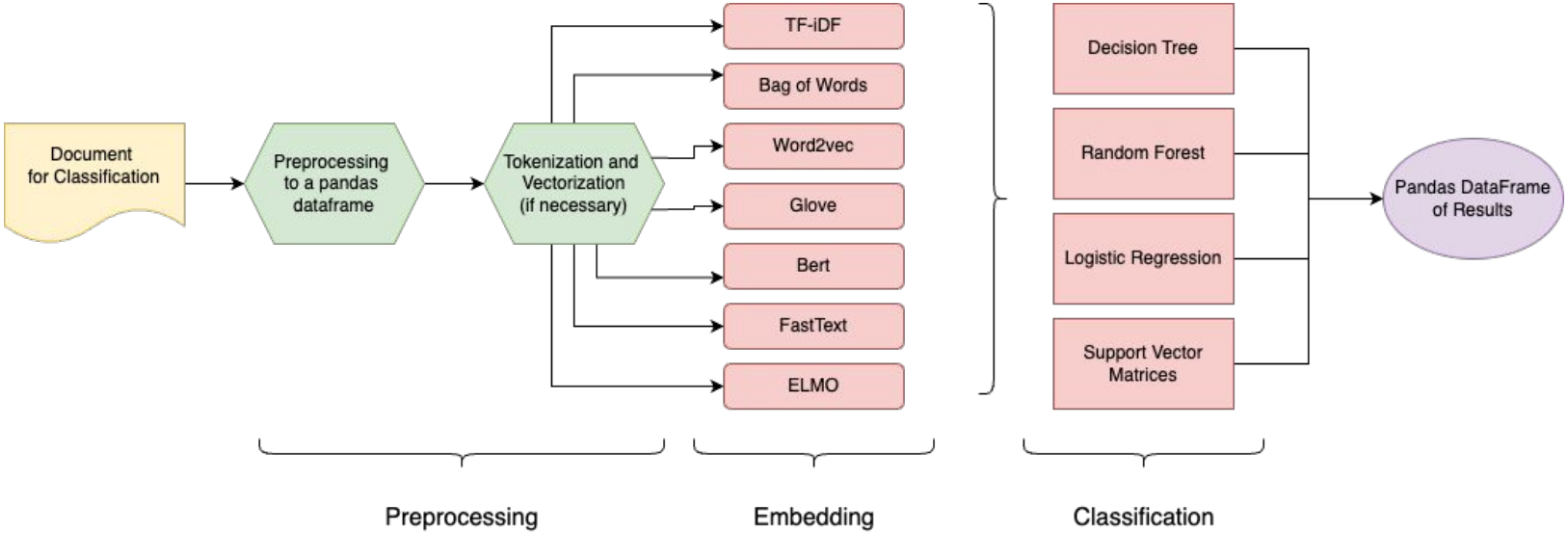
Contextual word embeddings that capture word meaning based on their surrounding context

Metrics

- Accuracy: Measures the overall correctness of the model's predictions by calculating the ratio of correctly classified instances to the total number of instances.
- Precision: Evaluates the proportion of true positive predictions out of all positive predictions.
- Recall: Assesses the proportion of true positive predictions out of all actual positive instances.
- F1 Score: Combines precision and recall into a single metric.
- **The selection algorithm allows the user to choose a specific metric based on their preferences or task requirements.**



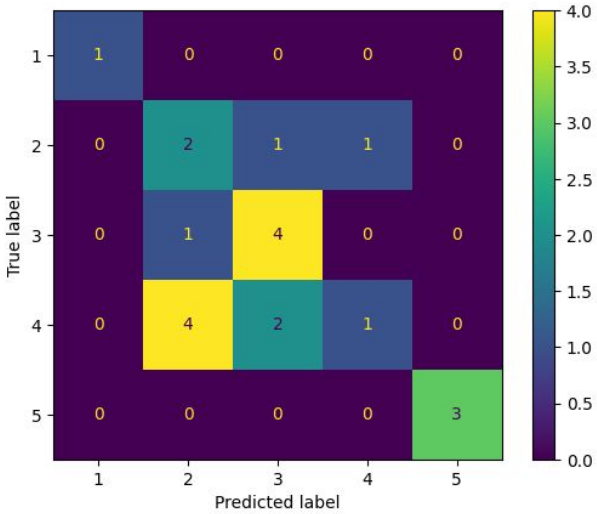
Component Diagram



Result Summary

Model	IMDB data (n=1000 samples)				Dow PSE data			
	RF	DT	LR	SVM	RF	DT	LR	SVM
bag-of-words	0.81	0.67	**0.83**	0.73	**0.80**	0.77	0.77	0.55
tf-idf	0.81	0.60	**0.83**	**0.83**	0.72	**0.73**	0.68	0.62
Word2Vec	0.75	0.56	**0.81**	0.78	**0.68**	0.58	0.67	0.53
fastText	**0.76**	0.59	0.58	0.67	**0.75**	0.47	0.45	0.45
ELMo	0.75	0.65	**0.82**	0.81	**0.75**	0.64	0.68	0.45
Bert	0.67	0.60	**0.82**	0.75	**0.77**	0.55	0.73	0.70
GloVe	**0.71**	0.56	**0.71**	0.65	**0.75**	0.55	0.53	0.27

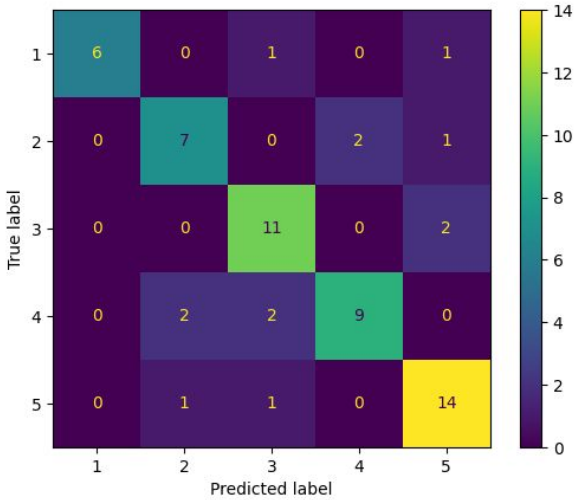
Results



ELMo Embeddings on same person reports, RF score 0.55

- As shown by this confusion matrix, NLP models are able to almost perfectly predict level 1 and 5 incidents for small datasets
- However, even contextual models struggle with determining the incident level on intermediate accidents

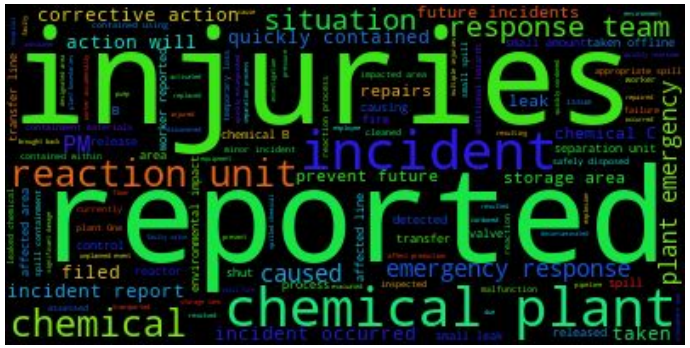
Results



ELMo Embeddings on Concatenated data Reports, RF score 0.783

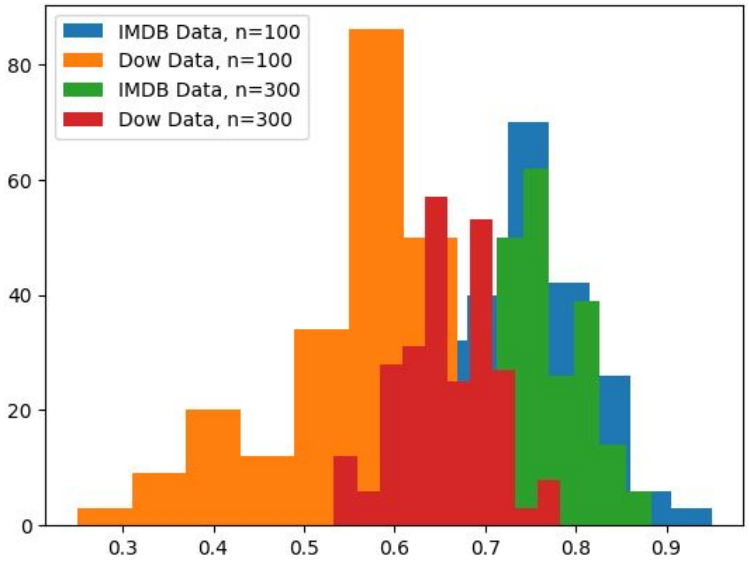
- As shown by this confusion matrix, NLP models are able to almost perfectly predict level 1 and 5 incidents for small datasets
- However, even contextual models struggle with determining the incident level on intermediate accidents.
- Increasing training data raises accuracy overall, however reduces the models ability to predict level 1 and 5 incidents

Analysis of Dow Data



- Lots of words repeatedly showing up
- Some entries are exactly the same except the chemical name
- Upon training on individual data sheet, Same Person Report has the best score, the other two sheets together has the worst score, and score of whole dataset in between

Analysis of Dow Data



- Trial with data preprocessing by removal of date and time, punctuations, a list of stop words and converting all words to lowercase
- Preprocessing not showing significant improvement over model performance
- High variance of training results, and random state can also affect the score a lot
- Implemented hyperparameter optimization including grid search, random search and Optuna, but data is easily overfitted due to the small sample size

Summary

- We built a package that automatically selects the best word embedding model from seven models for a specific downstream NLP task with a metrics defined by user
- Each model is tested with both IMDB and Dow dataset
- The selection algorithm outputs the model chosen, embeddings and a classifier which performs best with the embeddings
- For small datasets (< 1000 samples) low level embedding methods are best for classification tasks
- Simpler linear ML models tend to perform better than more complex ones
- Fine tuned models in tensorflow and pytorch are needed to get the best results from contextual embeddings

Acknowledgement

- Dr. Ivan Castillo, Dow Inc.
- Dr. You Peng, Dow Inc.
- Prof. Dave Beck
- Evan Komp, PhC

eScience Institute