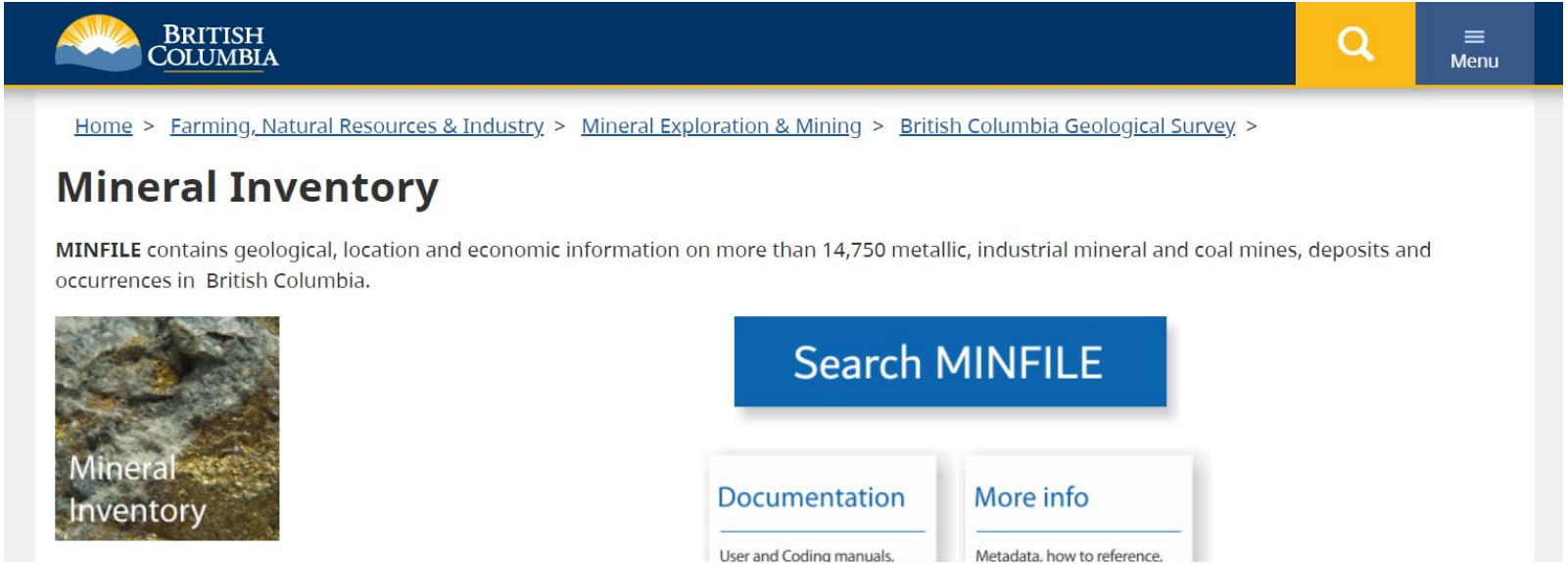
Capstone Project

Partnering with Minerva Intelligence

Jon Chan, Millie Lou, Ruby Nguyen, Zhongyu Pan

Background Information

- **Minerva Intelligence**: knowledge engineering in earth-sciences domain
- **The problem**: information extraction from geology reports
- **Objective**: provide a system for information extraction



Sample Minfile Report

SUMMARY Summary Help

Name	WAVERLEY (L.3597), M GUE (L.35 WAVERLEY-TANGIER	**NMI**	082N5 Ag2
		Mining Division	Revelstoke
Status	Developed Prospect	**BCGS Map**	082N041
Latitude	051° 27' 00"	**NTS Map**	082N05W
Longitude	117° 57' 35"	**UTM**	11 (NAD 83)
		Northing	5700305
		Easting	433311
Commoditie		**Deposit Types**	I05 : Polymetallic veins Ag-Pb-Zn+/-Au J01 : Polymetallic manto Ag-Pb-Zn
Tectonic Belt		**Terrane**	Kootenay, Ancestral North America

Capsule Geology

The Waverley occurrence is located on the western slope of Sorcerer Mountain, 1.75 kilometres southeast of the confluence of Sorcerer and Holway creeks and approximately 52 kilometres northeast of Revelstoke. The Tangier occurrence (MINFILE 082N 015) lies 750 metres to the west.

Average assays for samples taken from the main oreshoot on the Waverley claim were 5.8 per cent lead and 606.7 grams per tonne silver over an average width of 2 metres and a length of approximately 21 metres (Special Bulletin (1928), Report on Waverley-Tangier Property, by J.D. Galloway). A sample of ore from the No. 2 tunnel assayed 4.1 grams per tonne gold, 1588.1 grams per tonne silver, 2.1 per cent lead, 26.7 per cent zinc and 1.35 per cent copper (Geological Survey of Canada Summary Report 1928 Part A, page 179).

The ore occurs in well-defined fissures, replaces dark grey or black fine-grained limestone and is found as irregular bodies more or less elongated along predominant shear and fault zones that trend approximately 320 degrees. Veins of quartz and calcite, striking more northerly than the main oreshoots, are barren in most places.

The ore is highly oxidized and consists of limonite, anglesite, cerussite, malachite, azurite, smithsonite and occasional nodules of galena and tetrahedrite in a gangue of decomposed limestone, calcite and quartz.

Vocabulary:

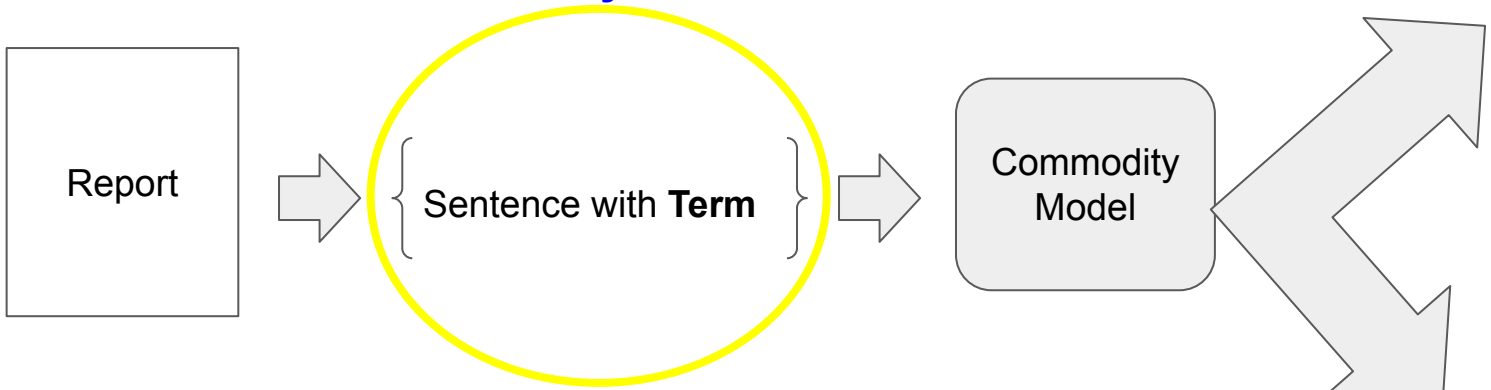
- Field
- Term
- Mention

Fields:

1. Commodity
2. Significant Mineral
3. Alteration Mineral
4. Deposit Character
5. Dominant Host Rock

General Approach

- **Task:** to build **5 binary classification models**



True
(Term is Commodity)

False
(Term is not Commodity)

- **Two types of models:**
  1. Feature-based models
  2. Neural-based models
- **Evaluation metric:** precision vs. recall

Build a Lexicon for Each Field

SUMMARY

Summary Help

Name	WAVERLEY (L.3597), MONTAGUE (L.3596), WAVERLEY-TANGIER	**NMI**	082N5 Ag2
		Mining Division	Revelstoke
		BCGS Map	082N041
Status	Developed Prospect	**NTS Map**	082N05W
Latitude	051º 27' 00"	**UTM**	11 (NAD 83)
Longitude	117º 57' 35"	**Northing**	5700305
		Easting	433311
Commodities	Lead, Silver, Gold, Zinc, Copper	**Deposit Types**	I05 : Polymetallic veins Ag-Pb-Zn+/-Au J01 : Polymetallic manto Ag-Pb-Zn
Tectonic Belt	Omineca	**Terrane**	Kootenay, Ancestral North America

Capsule Geology

The Waverley occurrence is located on the western slope of Sorcerer Mountain, 1.75 kilometres southeast of the confluence of Sorcerer and Holway creeks and approximately 52 kilometres northeast of Revelstoke. The Tangier occurrence (MINFILE 082N 015) lies 750 metres to the west.

Average assays for samples taken from the main oreshoot on the Waverley claim were 5.8 per cent lead and 606.7 grams per tonne silver over an average width of 2 metres and a length of approximately 21 metres (Special Bulletin (1928), Report on Waverley-Tangier Property, by J.D. Galloway). A sample of ore from the No. 2 tunnel assayed 4.1 grams per tonne gold, 1588.1 grams per tonne silver, 2.1 per cent lead, 26.7 per cent zinc and 1.35 per cent copper (Geological Survey of Canada Summary Report 1928 Part A, page 179).

The ore occurs in well-defined fissures, replaces dark grey or black fine-grained limestone and is found as irregular bodies more or less elongated along predominant shear and fault zones that trend approximately 320 degrees. Veins of quartz and calcite, striking more northerly than the main oreshoots, are barren in most places.

The ore is highly oxidized and consists of limonite, anglesite, cerussite, malachite, azurite, smithsonite and occasional nodules of galena and tetrahedrite in a gangue of decomposed limestone, calcite and quartz.

Lexicon:

- Term = Lead
- Mentions = {**'lead'**}

+

Mineral Taxonomy

=

{'lead', **'anglesite', 'apatite', 'cerussite', 'galena', 'wulfenite'**}

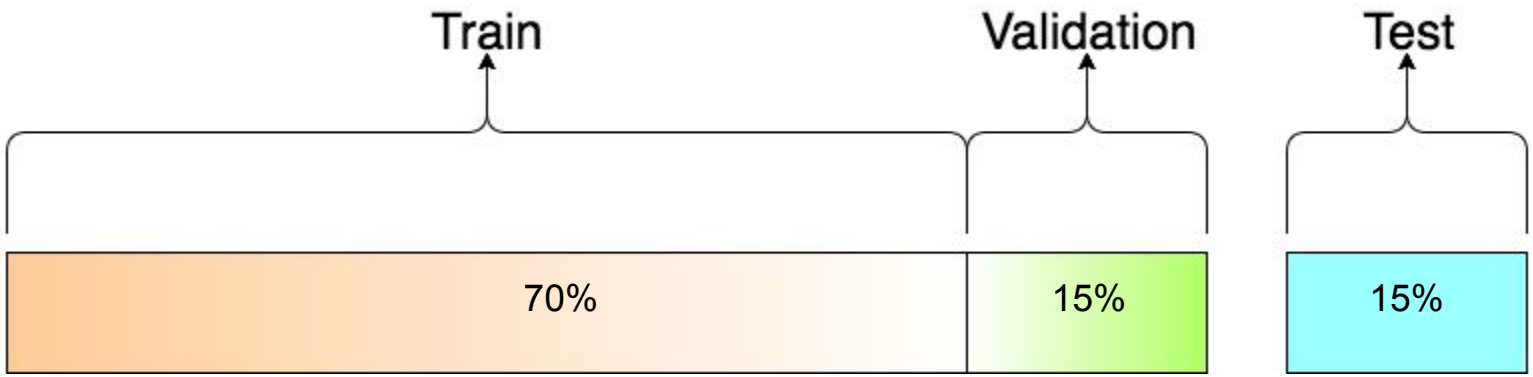
+

Morphology

=

{'lead', 'anglesite', 'apatite', 'cerussite', 'galena', 'wulfenite', **'leaded', 'leaden'**}

Split Datasets



- 10,567 reports for training models
- 2,264 reports for evaluating models during model development
- 2,265 reports for a final model evaluation at the end of the project

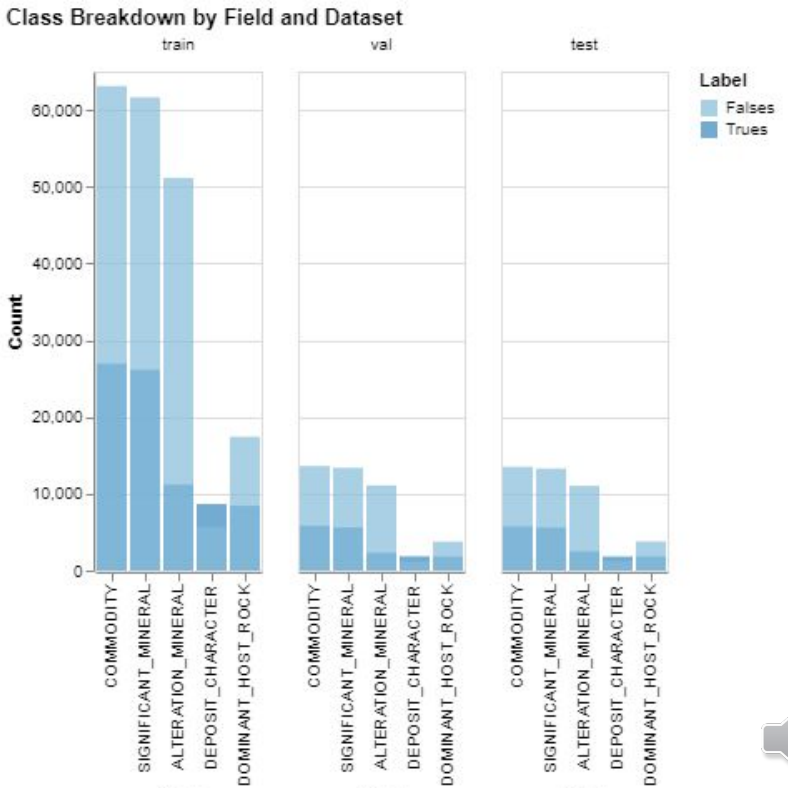
Image: https://towardsdatascience.com/train-validation-and-test-sets-72cb40cba9e7

Compile Sentences with Mentions and Label Terms

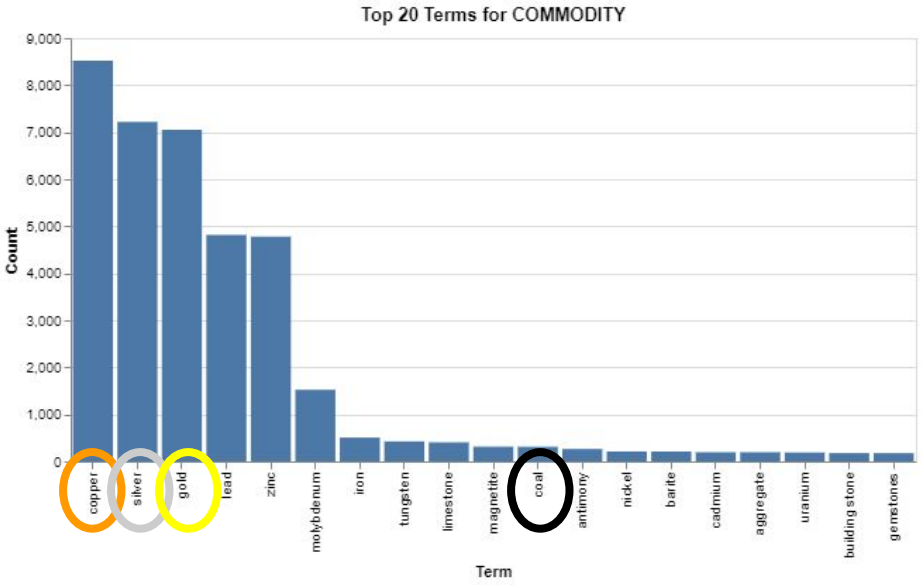
	MINFILNO	Term	Sentences	Original_Sents	Term_ID_Feature	Is_Labeled
0	092GSW015	germanium	[[it, is, a, fairly, plastic, clay, with, some...	[[It, is, a, fairly, plastic, clay, with, some...	[[0, 0, 0, 0, 0, 1, 0, 0, 0, 0]]	False
1	092GSW015	clay	[[it, is, a, fairly, plastic, clay, with, some...	[[It, is, a, fairly, plastic, clay, with, some...	[[0, 0, 0, 0, 0, 1, 0, 0, 0, 0]]	True
2	092L 366	aggregate	[[the, lot, 4, aggregate, occurrence, is, loca...	[[The, Lot, 4, aggregate, occurrence, is, loca...	[[0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]...	True
3	082GNW079	quartzite	[[the, area, is, underlain, by, mid, proterozo...	[[The, area, is, underlain, by, mid, Proterozo...	[[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...	False
4	082GNW079	argillite	[[the, area, is, underlain, by, mid, proterozo...	[[The, area, is, underlain, by, mid, Proterozo...	[[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...	False

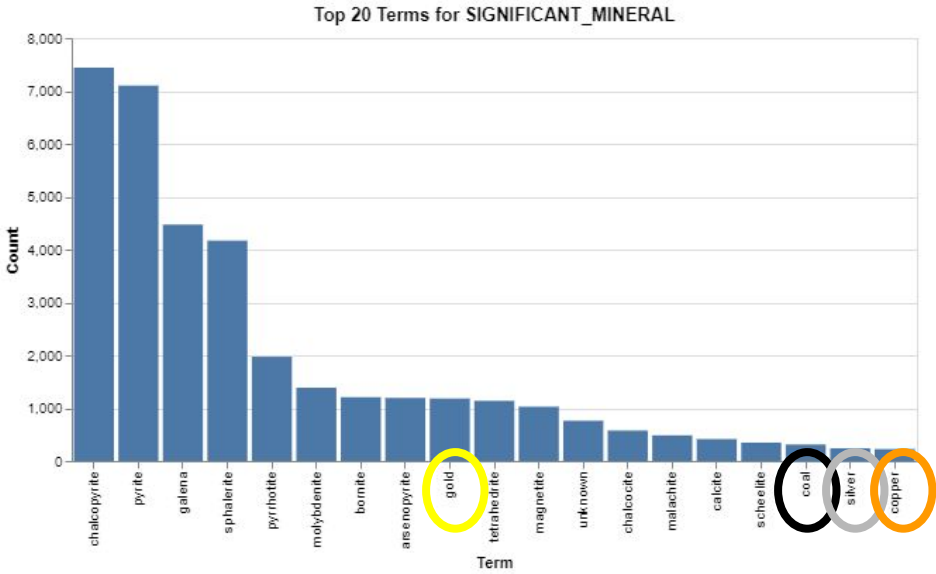
Understanding the Data

Field	No. of Terms
Commodity	152
Significant Mineral	371
Alteration Mineral	195
Deposit Character	15
Dominant Host Rock	8



Understanding the Data





Understanding the Data

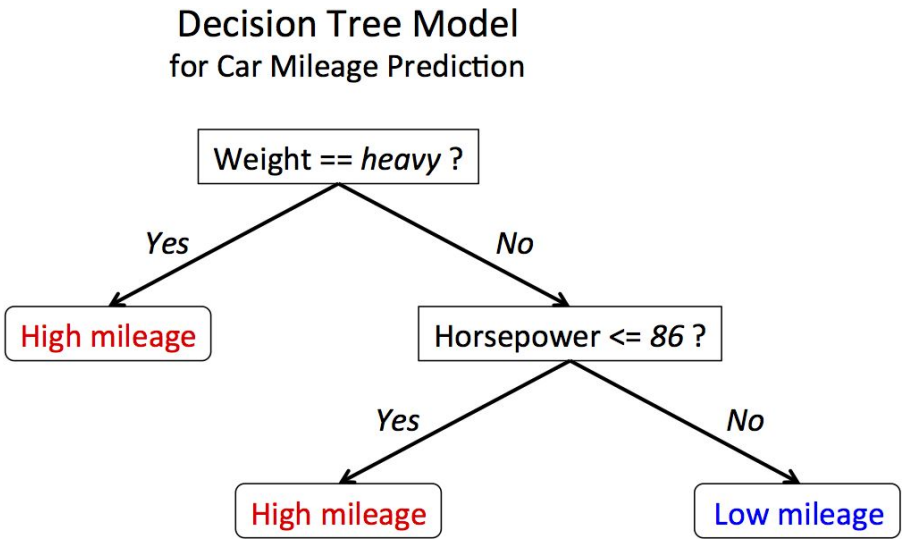
Unmatched Terms

Field	Train (count, % of dataset)		Validation (count, % of dataset)		Test (count, % of dataset)	
Commodity	1,558	5.46%	336	5.47%	321	5.3%
Significant Mineral	2,125	7.52%	475	7.8%	446	7.39%
Alteration Mineral	7,360	39.64%	1575	40.43%	1530	37.98%
Deposit Character	6,012	40.97%	1242	40.16%	1284	41.37%
Dominant Host Rock	1,625	16.16%	346	16.06%	343	15.89%

Feature-based Model

LightGBM Classifier (LGBM)

- A gradient boosting model
- Tree based
- One for each field



Feature Set Generation

Example sentence for commodity field:

Fractures contain electrum with low gold values.

Term	gold
Mention In Text	electrum
Previous Word	contain
Previous 2 Words	fractures contain
Next Word	with
Next 2 Words	With low
Position	3

Term	silver
Mention In Text	electrum
Previous Word	contain
Previous 2 Words	fractures contain
Next Word	with
Next 2 Words	with low
Position	3

Term	gold
Mention In Text	Electrum, gold
Previous Word	contain, low
Previous 2 Words	fractures contain, with low
Next Word	with, values
Next 2 Words	with low, values <BD>
Position	3, 6

- Feature sets are then converted into a vector representation

Feature-based Model Tuning

- Manual hyperparameter tuning

Field: Dominant Host Rock	Precision	Recall
LGBM Default	0.7681	0.5950
LGBM Manually Tuned	0.7805	0.6132

- Optuna

Field: Commodity	Precision	Recall
LGBM Default	0.9195	0.9302
LGBM Optuna	0.9280	0.9393

The above results were obtained from passing the test dataset into the models

Linguistic Features

10 features

- Term
- Mention
- N-gram
- Mention average position
- Mention count.
- Bag-of-words

```
{'term_argillite': True,
 'mention_argillites': True,
 'prev_unigram_and': True,
 'prev_bigram_greywackes_and': True,
 'next_unigram_.': True,
 'next_bigram_._<BD>': True,
 'mentions_count': 2,
 'avg_position': 0.4375,
 'word_b4_mention_<BD>': True,
 'word_b4_mention_the': True,
 'word_b4_mention_area': True,
 'word_b4_mention_is': True,
 'word_b4_mention_underlain': True,
 'word_b4_mention_by': True,
 'word_b4_mention_mid': True,
```

Linguistic Features

- Capital mention (e.g Gold Coast)

one sample (rc 5) taken at this site assayed 2.04 per cent zinc , 0.16 per cent lead , 15.7 grams per tonne **silver** and 0.31 grams per tonne gold (assessment report 17670) .
a strong northwest fault cuts diagonally across the skarn and localizes massive **chalcopyrite** with pods of chalcocite and **bornite** .
the company reports that about 7645 cubic metres (10,000 cubic yards) of material were processed and 11,952 grams (421.6 ounces) of gold were recovered and 1134 grams (40
sample w 1 assayed 0.265 grams per tonne gold , 13.3 grams per tonne **silver** , 1.973 per cent copper and 11.069 per cent iron (assessment report 16860) .
an average sample from the deposit found at the higher elevation , taken across 3.7 metres , assayed 0.8 per cent copper , 13.71 grams per tonne **silver** and a trace of gold (min

- Unit and amount : {unit_grams_per_tonne: 15.7}
  - Increase the precision

Feature Importances

- An attempt to understand whether certain features are more important
- Features are split into four categories - Term, mention, n-grams, bag-of-words

Positive bag-of-words features (Commodity)	'word_b4_mention_assayed', **'word_b4_mention_tonnes'**, 'word_b4_mention_pyrite', **'word_b4_mention_grams'**, 'word_after_mention_occur'
Positive n-grams (Commodity)	**'prev_unigram_tonne'**, **'prev_bigram_per_tonne'**, **'prev_bigram_per_cent'**, **'prev_unigram_cent'**, 'next_unigram_mineralization'
Negative n-grams (Significant Minerals)	'prev_unigram_quartz', 'next_unigram_veins', **'prev_bigram_per_tonne'**, **'prev_unigram_tonne'**, 'next_unigram_vein'
Positive n-grams (Alteration Minerals)	**'next_unigram_altered'**, **'next_bigram_carbonate_alteration'**, 'next_unigram_sericite', 'prev_unigram_sericite', 'next_unigram_zone'

Neural-based model

CNN Diagram

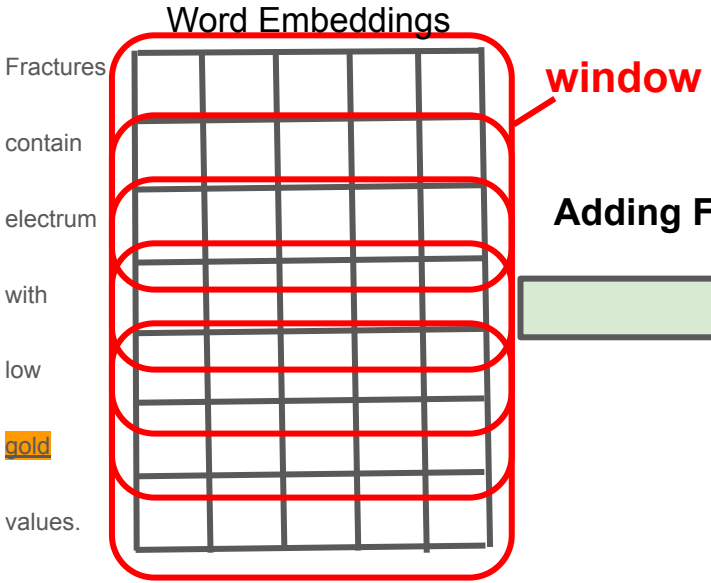
Word Embeddings

CNN Experiments

General Ideas

- CNN only takes in numbers.
- We need to convert data from text to numbers representation.
- CNN learns the patterns between numbers representation and labels.

CNN Diagram



Adding Features

Mentions Distance Feature

[5, 4, 3, 2, 1, 0, 1]

['Fractures','contain','electrum','with','low','gold','values']

Term Binary Feature

[0, 0, 1, ..., 0]

["silver", "copper", "gold", ..., "zinc"]

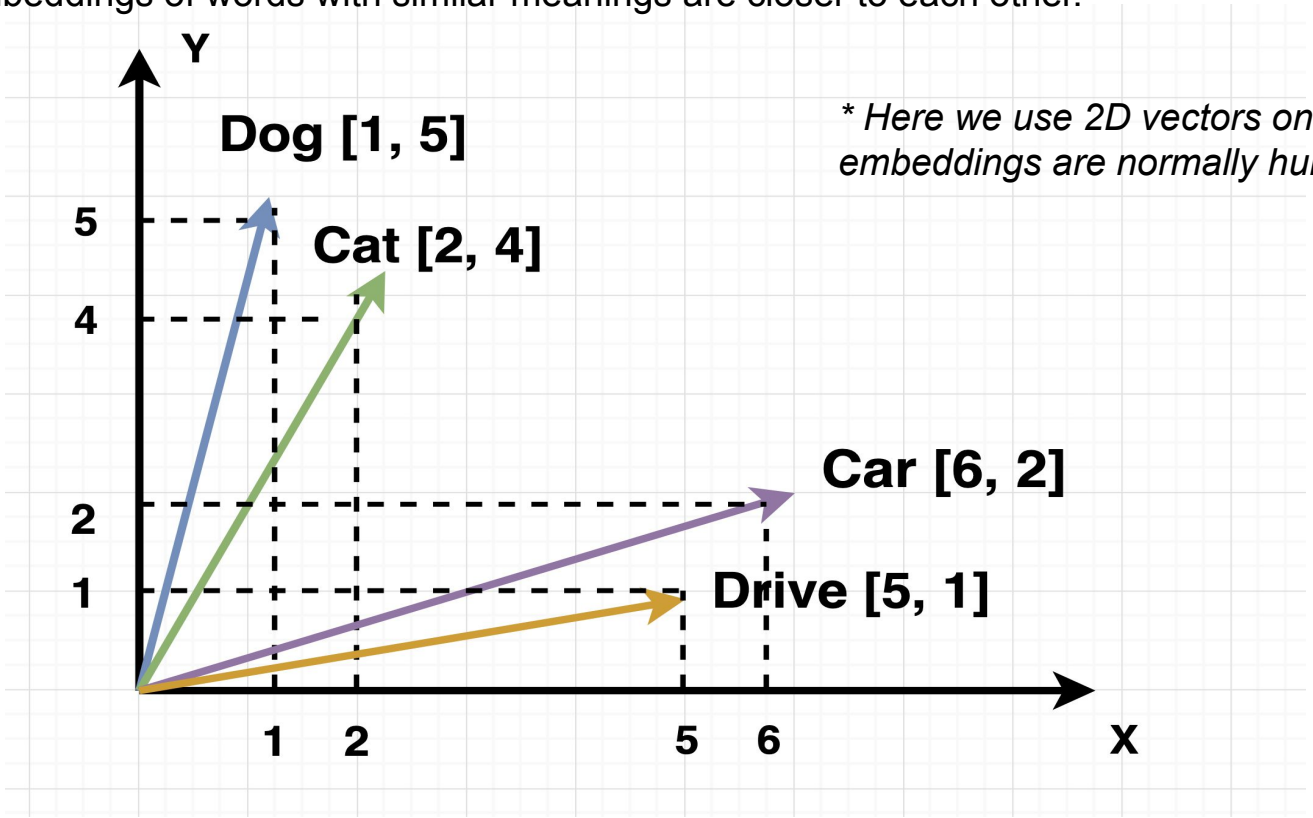
Other linguistics feature (Future work)

Using words and features information to **classify.**

Classes	Labels
Term is Commodity	1
Term is not Commodity	0

Word Embeddings

- An embedding is just a vector,
- Embeddings of words with similar meanings are closer to each other.



** Here we use 2D vectors only for simplicity, embeddings are normally hundreds of dimensions.*

CNN Experiments

* (Precision ,Recall) of Test Set for Commodity Field

CNN + Mentions Distance + Term Binary Feature (0.91, 0.93)

CNN + Mentions Distance Feature (0.90, 0.91)

CNN + Pre-trained Word Embeddings (0.78, 0.77)

Baseline CNN (0.76, 0.79)

Dummy Classifier (0.29, 0.29)

Final Evaluation and Analysis

	Validation Dataset				Test Dataset			
	Best LGBM		Best CNN		Best LGBM		Best CNN	
	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall
Commodity	0.9251	0.9334	0.9119	0.9239	0.9280	0.9393	0.9132	0.9264
Significant Mineral	0.8740	0.9157	0.8017	0.9583	0.8842	0.9174	0.8085	0.9569
Alteration Mineral	0.7349	0.6902	0.6197	0.7062	0.7688	0.7162	0.6684	0.8078
Deposit Character	0.8233	0.8455	0.8118	0.8671	0.7991	0.8632	0.7767	0.8511
Dominant Host Rock	0.7788	0.6208	0.7118	0.6882	0.7805	0.6132	0.7344	0.6733

Note: Precision and Recall are for positive class only.

Future Works

- Deeper error analysis
- Expand lexicon
- Implement feature selection for individual feature model
- Deeper analysis for negation
- Using BERT
- Pre-train its own word embeddings for the geology domain
- Add features from feature-based model to neural-based model

Thank You!

Clinton Smyth

Anna Hicken

Julian Brooke