A.4: Final Report

Gurjus Singh

MSDS 453: Natural Language Processing Sec 56

Professor Jennifer Sleeman

Northwestern University

Introduction & Problem Statement

In this research, the goal is to use Deep Learning to do binary classification of whether tweets are in English. Deep Learning was used in particular as in past research there had already been experiments done with Machine Learning Models and particular Deep Learning has a benefit of automatically learning from it's own errors through Back Propagation (Grossfeld, 2021, n.pg.). The classification will be based on a dataset on Kaggle which collected 10,502 tweets from Twitter which were in various languages (Tatman, 2017, n.pg.). The column to classify on was decided based on the Ontology Analysis research done in the past. The construction of ontology was revised after the last research and a 4th iteration of the ontology was created.

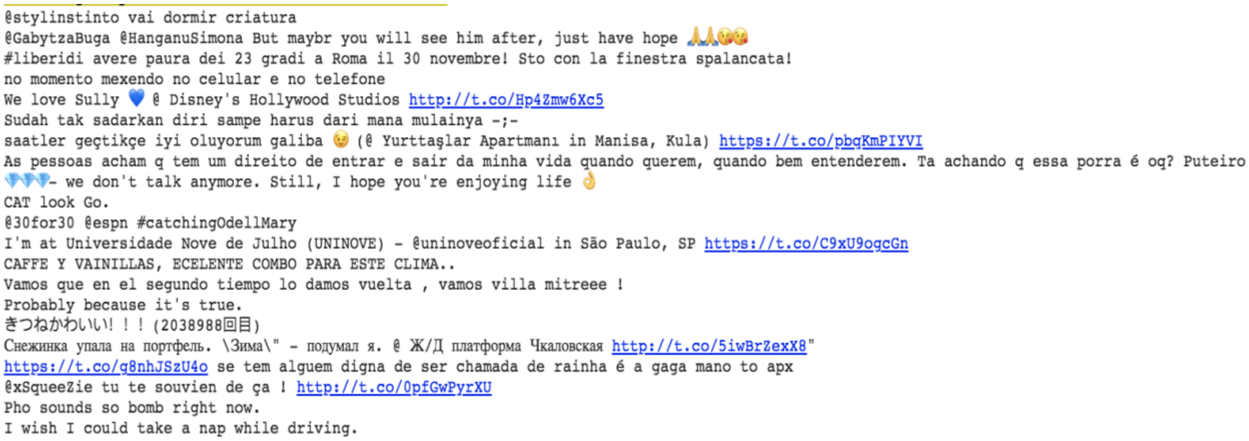
Literature Review

One related article to this research that was found on the web was titled "Deep Neural Network Language Identification". The author first mentions that the dataset is large with 328 distinct languages and around 6 million sentences (O'Sullivan,2020, n.pg.). In the research, a trigram was used for data preprocessing of the language text (O'Sullivan,2020, n.pg.). After trigrams were created, then they author identified the 200 most common trigrams in the language, and found the most unique ones (O'Sullivan,2020, n.pg.). After the trigrams were found and extracted, CountVectorizer in Python was used to vectorize the text (O'Sullivan,2020, n.pg.). Lastly, after vectorizing, the features were scaled before training the Neural Network model (O'Sullivan,2020, n.pg.). The Neural Network architecture used in the author's research was a DNN architecture and resulted in a 98.26 percent test accuracy (O'Sullivan,2020, n.pg.).

The author’s research had some related methods that were applied to this research. For example, the author used DNN architecture similarly to what was used in this research and, the author found a dataset that contained text in different languages, which was also similar to this research’s dataset.

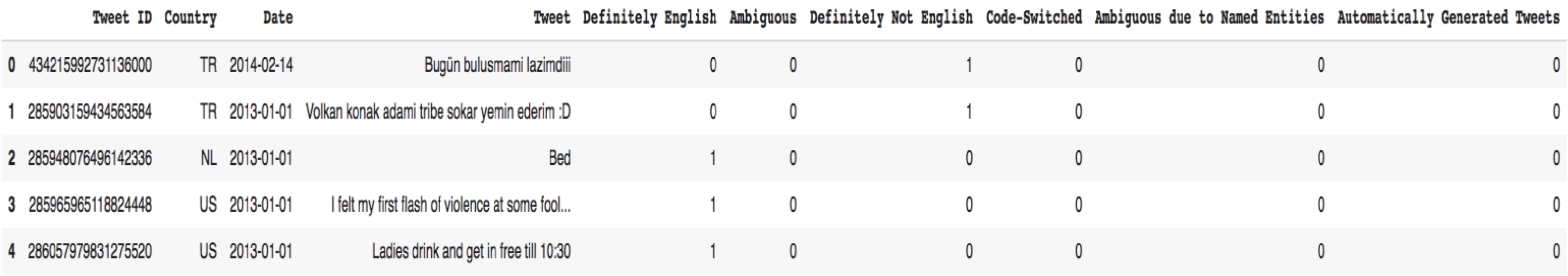
Data

The data used in binary classification was taken from Kaggle, which consisted of 10,502 tweets written in world languages. The columns of the dataset included attributes of tweet such as where it was geotagged, “date written” , whether it was “English or not”, whether the tweet was “Automatically Generated”, whether the tweet was “Ambiguous” to tell which language it was composed of and whether “Code Switching” was involved. Codeswitching is when a person uses two languages in a single conversation (Akslov, 2016, n.pg.). The columns “English or not”, “Automatically Generated”, “Ambiguous”, “Code Switching” have all been one hot encoded with ones and zeroes as seen in 1-2. In 1-1, there is an initial look at the data, and in 1-3, there are the tweets classified to their exact language with the spacy package, which gives the language of the tweet.



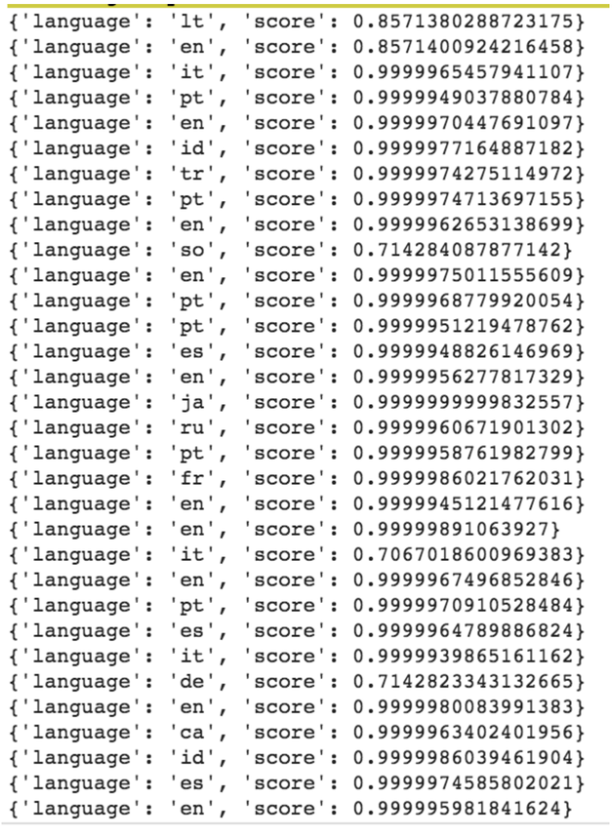
```
@stylinstinto vai dormir criatura
@GabytzaBuga @HanganuSimona But maybr you will see him after, just have hope 🙏🙏😘😘
#liberidi avere paura dei 23 gradi a Roma il 30 novembre! Sto con la finestra spalancata!
no momento mexendo no celular e no telefone
We love Sully 💙 @ Disney's Hollywood Studios http://t.co/Hp4Zmw6Xc5
Sudah tak sadarkan diri sampe harus dari mana mulainya -;-
saatler geçtikçe iyi oluyorum galiba 😉 (@ Yurttaşlar Apartmanı in Manisa, Kula) https://t.co/pbqKmPIYVI
As pessoas acham q tem um direito de entrar e sair da minha vida quando querem, quando bem entenderem. Ta achando q essa porra é oq? Puteiro
💎💎💎- we don't talk anymore. Still, I hope you're enjoying life 👌
CAT look Go.
@30for30 @espn #catchingOdellMary
I'm at Universidade Nove de Julho (UNINOVE) - @uninoveoficial in São Paulo, SP https://t.co/C9xU9ogcGn
CAFFE Y VAINILLAS, ECELENTE COMBO PARA ESTE CLIMA..
Vamos que en el segundo tiempo lo damos vuelta , vamos villa mitreee !
Probably because it's true.
きつねかわいい! ! ! (2038988回目)
Снежинка упала на портфель. \Зима\" - подумал я. @ Ж/Д платформа Чкаловская http://t.co/5iwBrZexX8"
https://t.co/g8nhJSzU4o se tem alguem digna de ser chamada de rainha é a gaga mano to apx
@xSqueeZie tu te souvien de ça ! http://t.co/0pfGwPyrXU
Pho sounds so bomb right now.
I wish I could take a nap while driving.
```

Initial Look at Dataset; gives a deeper understanding of what formatting of tweets looks like such as syntax, emojis, and also identifying the languages 1-1



	Tweet ID	Country	Date	Tweet	Definitely English	Ambiguous	Definitely Not English	Code-Switched	Ambiguous due to Named Entities	Automatically Generated Tweets
0	434215992731136000	TR	2014-02-14	Bugün bulusmami lazimdiii	0	0	1	0	0	0
1	285903159434563584	TR	2013-01-01	Volkan konak adami tribe sokar yemin ederim :D	0	0	1	0	0	0
2	285948076496142336	NL	2013-01-01	Bed	1	0	0	0	0	0
3	285965965118824448	US	2013-01-01	I felt my first flash of violence at some fool...	1	0	0	0	0	0
4	286057979831275520	US	2013-01-01	Ladies drink and get in free till 10:30	1	0	0	0	0	0

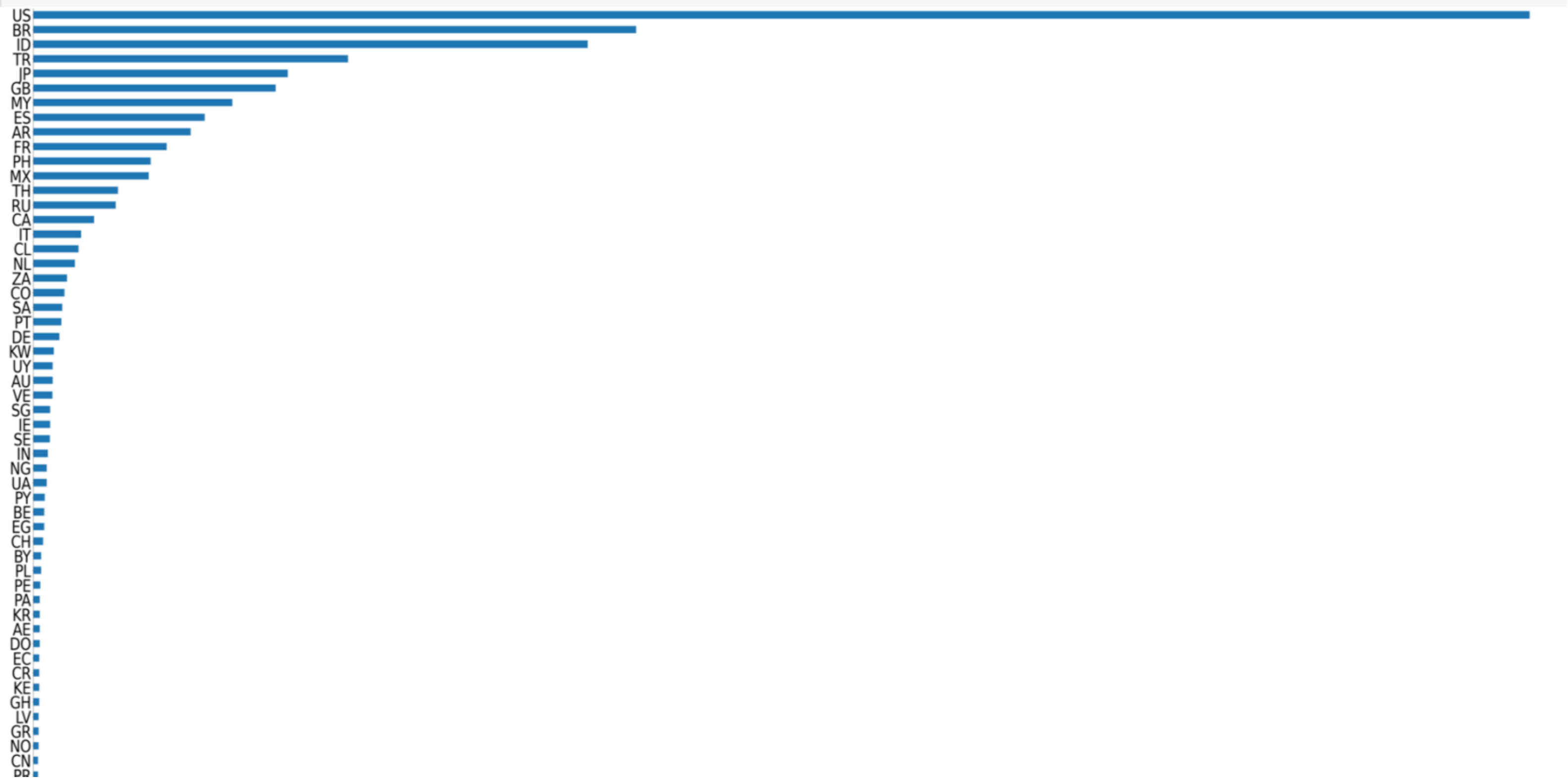
DataFrame of Language Dataset with Columns; gave an understanding of what metadata describes the tweet for example gives the country where tweet originated as well as if Tweet contains English 1-2



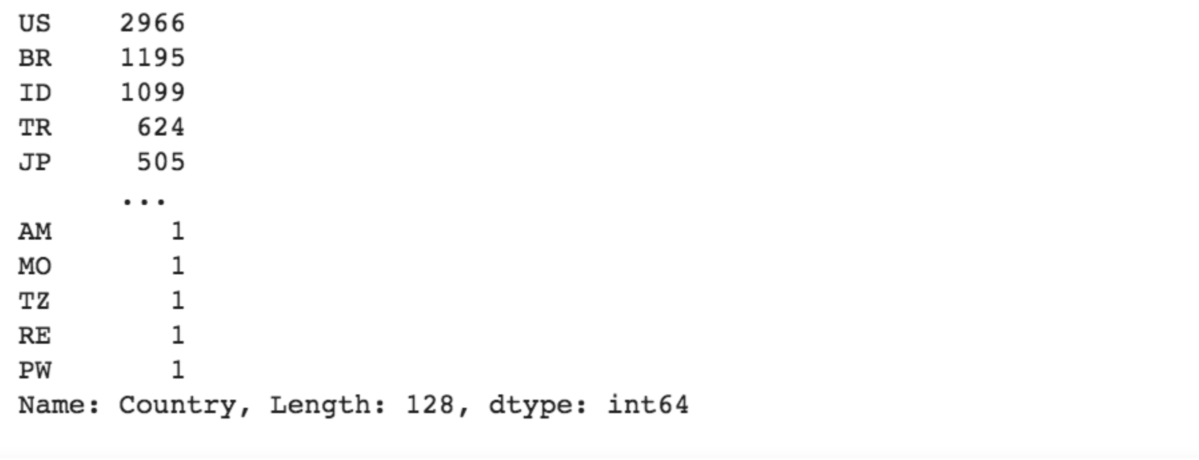
```
{'language': 'lt', 'score': 0.8571380288723175}
{'language': 'en', 'score': 0.8571400924216458}
{'language': 'it', 'score': 0.9999965457941107}
{'language': 'pt', 'score': 0.9999949037880784}
{'language': 'en', 'score': 0.9999970447691097}
{'language': 'id', 'score': 0.9999977164887182}
{'language': 'tr', 'score': 0.9999974275114972}
{'language': 'pt', 'score': 0.9999974713697155}
{'language': 'en', 'score': 0.9999962653138699}
{'language': 'so', 'score': 0.714284087877142}
{'language': 'en', 'score': 0.9999975011555609}
{'language': 'pt', 'score': 0.9999968779920054}
{'language': 'pt', 'score': 0.9999951219478762}
{'language': 'es', 'score': 0.9999948826146969}
{'language': 'en', 'score': 0.9999956277817329}
{'language': 'ja', 'score': 0.9999999999832557}
{'language': 'ru', 'score': 0.9999960671901302}
{'language': 'pt', 'score': 0.9999958761982799}
{'language': 'fr', 'score': 0.9999986021762031}
{'language': 'en', 'score': 0.9999945121477616}
{'language': 'en', 'score': 0.99999891063927}
{'language': 'it', 'score': 0.7067018600969383}
{'language': 'en', 'score': 0.9999967496852846}
{'language': 'pt', 'score': 0.9999970910528484}
{'language': 'es', 'score': 0.9999964789886824}
{'language': 'it', 'score': 0.9999939865161162}
{'language': 'de', 'score': 0.7142823343132665}
{'language': 'en', 'score': 0.9999980083991383}
{'language': 'ca', 'score': 0.9999963402401956}
{'language': 'id', 'score': 0.9999986039461904}
{'language': 'es', 'score': 0.9999974585802021}
{'language': 'en', 'score': 0.999995981841624}
```

Classification of Tweets Spacy package ; useful for further research if labels need to be added; languages were not identified in dataset; these labels could be used to train neural networks in future research

1-3



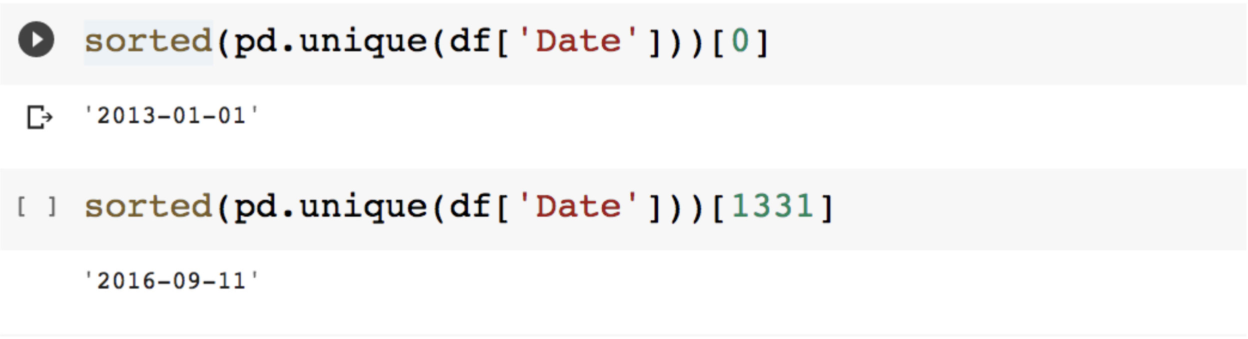
Countries where Tweets are from; get a deeper understanding of where tweets are from; US appears at top; Japan in 5th 1-4



```
US    2966
BR    1195
ID    1099
TR     624
JP     505
      ...
AM       1
MO       1
TZ       1
RE       1
PW       1
Name: Country, Length: 128, dtype: int64
```

Number of Tweets by Country; gives a numerical value of frequent tweets of a country; also gives a sense that multiple languages could be seen in one country 1-5

In 1-1, one can see that Spanish tweets are in the dataset; also, can see tweet structure such as emojis, and in 1-3, spacy also shows languages such as Japanese, Portuguese, Indonesian, and Spanish in the dataset; spacy can be important for labeling purposes in future research where multiclass classification on identification of languages can be used. In 1-4 one can see that countries where tweets are from are mostly from United States, Brazil, Indonesia, Turkey, Japan, and United Kingdom. On examining this further in 1-5, one can see that United States has 2966 Tweets out of 10,502, and Japan has 505 Tweets making it in the top 5. Further investigation of the dataset shows the Tweets were collected between the years 2013 to 2016 as seen in 1-6.



```
sorted(pd.unique(df['Date']))[0]
```

```
'2013-01-01'
```

```
[ ] sorted(pd.unique(df['Date']))[1331]
```

```
'2016-09-11'
```

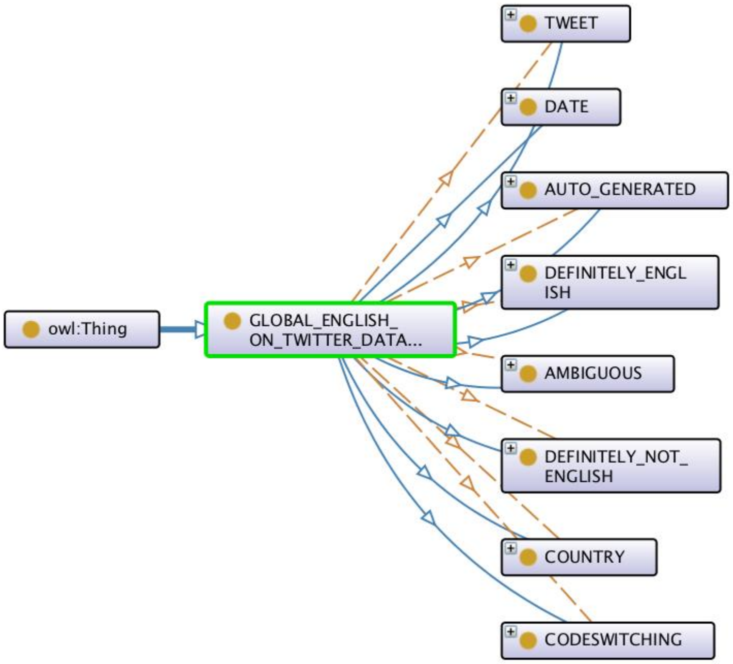
Dates of Tweets in Dataset; gets a deeper understanding of what kind of topics were trending on Twitter during these date ranges 1-6

Research Design and Modeling Method

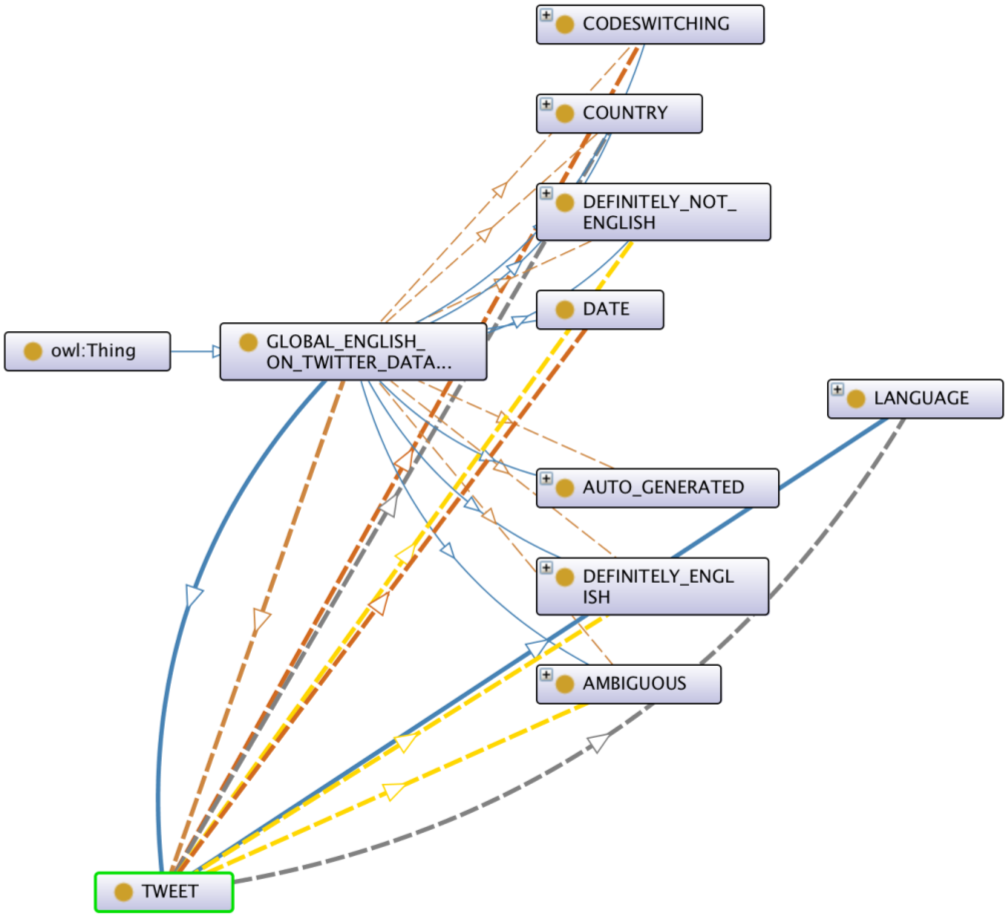
Python was used to do initial analysis of the dataset. A **for loop** in Python was used to print out the tweets, while the **Spacy** package was useful for identification of languages. Python was also used to clean the data such as removal of punctuation and stop words in English. **Protégé** was used to develop the ontology iterations. An ontology is defined as "specifications of the terms domain and the relations among them" (Noy and McGuiness, n.d, n.pg.).

For this particular research, different Deep Learning Architectures were used such as Deep Neural Networks, Recurrent Neural Networks and Long Short-Term Memory Neural Network. DNN is a type of Neural Network which is a Feed Forward Network meaning they only go in one direction and do not go backwards when computing data (SPRH LABS, 2019, n.pg) . **Relu** was used as the activation function for dense layers. The activation for output layer used was "**Sigmoid**". RNN is different from DNN in that it has internal memory. All the inputs are related to each other in this neural network unlike DNN (SPRH LABS, 2019, n.pg). Problems with this neural network is that there is a Gradient vanishing and exploding problems making it hard for models to learn and retain information (SPRH LABS, 2019, n.pg). LSTM resolves the vanishing gradient problem that RNN has (SPRH LABS, 2019, n.pg). Has three gates mainly input, forget, and output gates (Mittal, 2019, n.pg). Uses Time Series Data, and also does backpropagation (Mittal, 2019, n.pg). **Relu** was used as the activation function and **Sigmoid** for the output layer. Sigmoid is important for binary classification as the value ranges from 0 to 1 (Mittal, 2019, n.pg). Before using the models, the data was split on an 80-20 percent split between Train and Test set. There was also a 70-30 percent split between Train and Validation sets.

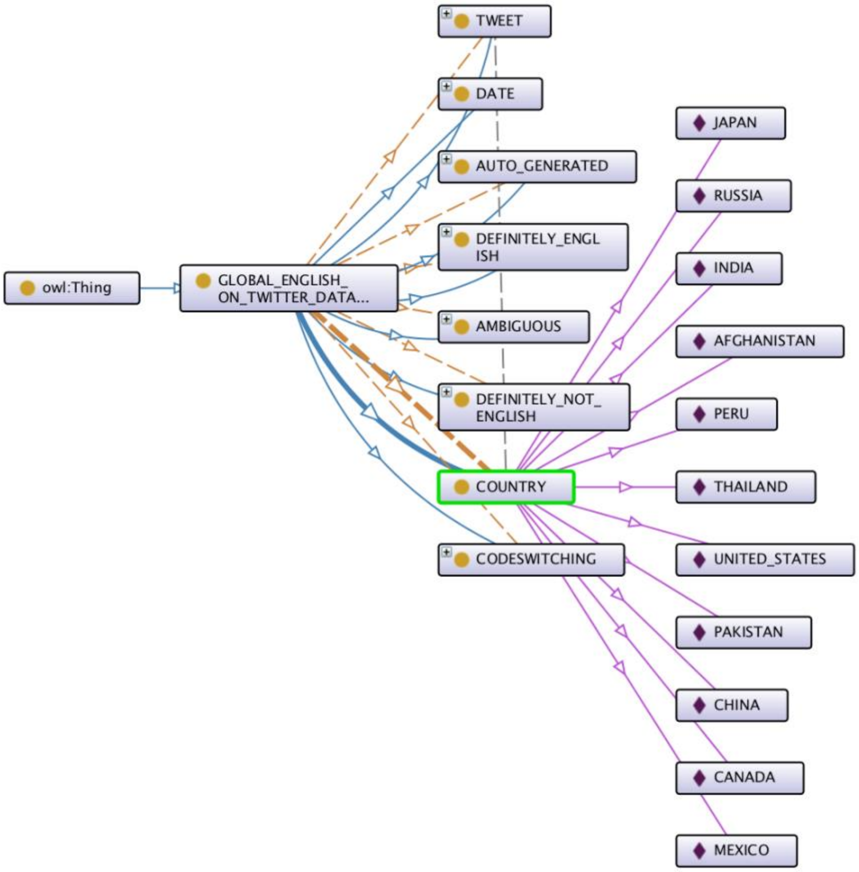
Results



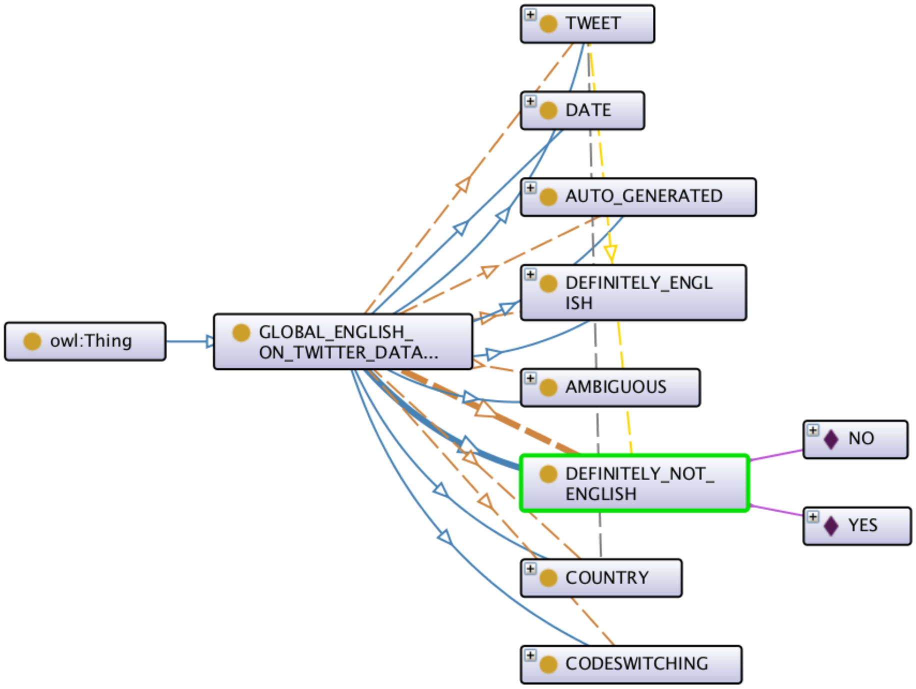
Ontology 4th Iteration shown in Parts; shows what the Dataset Contains such as column 1-7



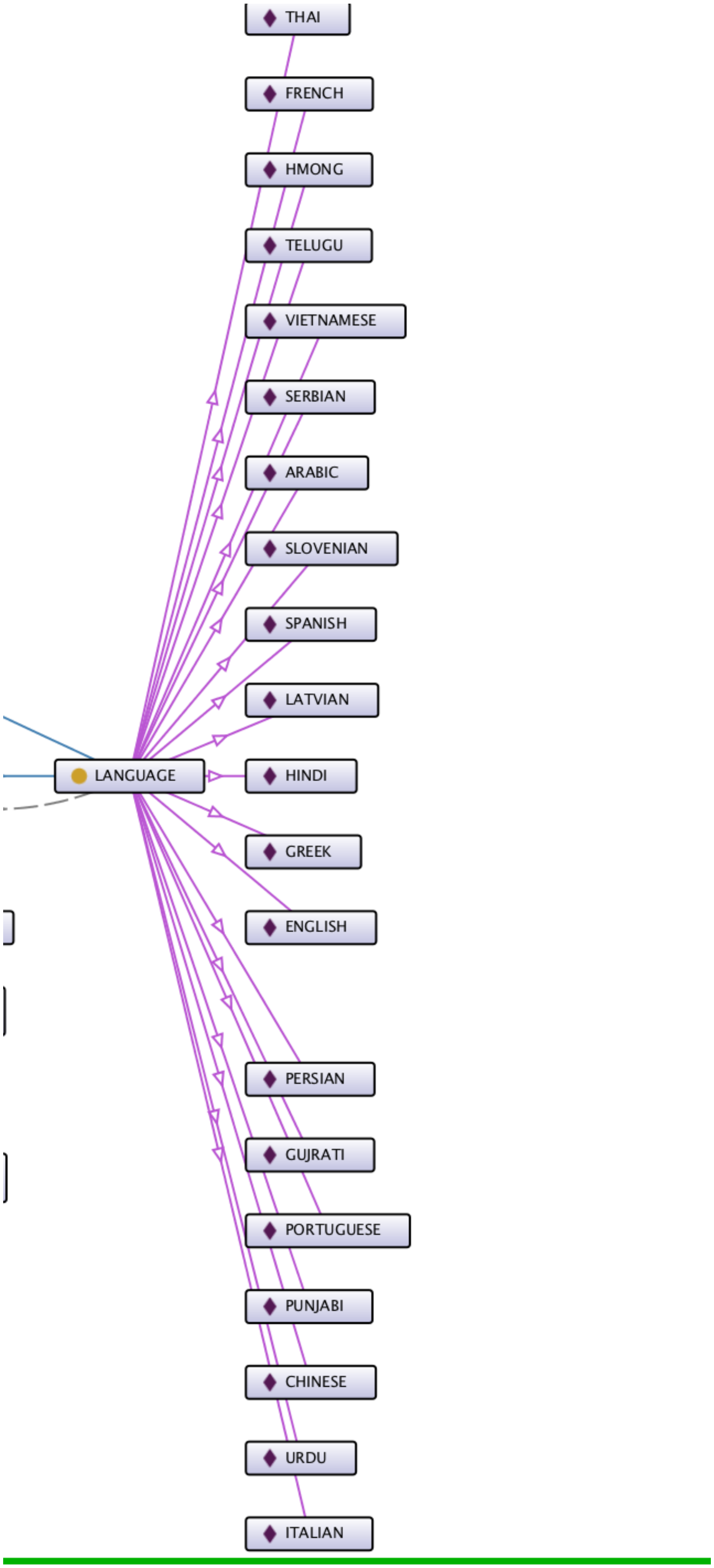
Ontology 4th Iteration shown in Parts; shows Tweet component of Dataset and object relationships such as Tweet has subclass called languages and Tweet has relationship with Country because Tweet has a country of origination 1-8



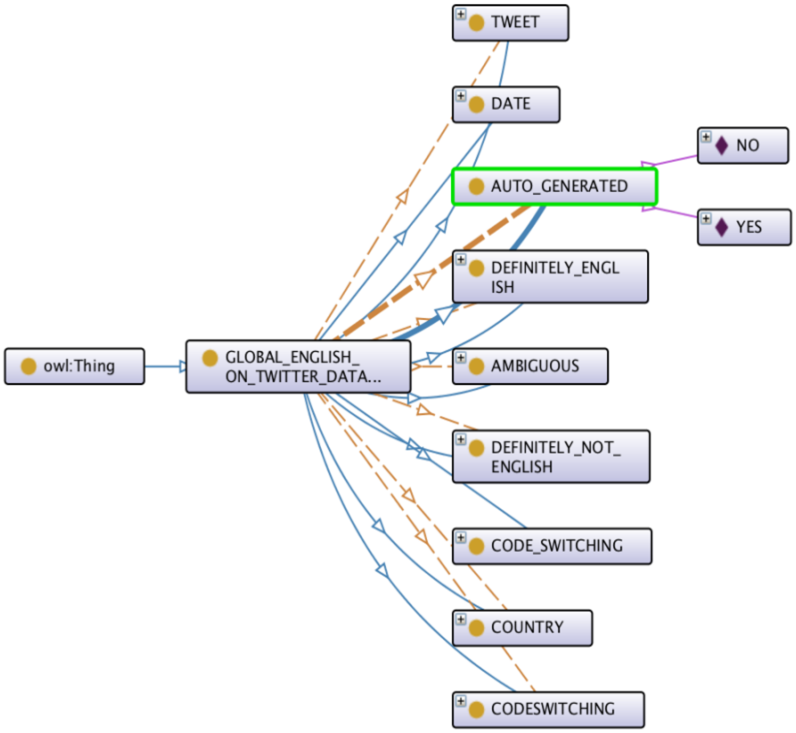
Ontology ^{4th} Iteration shown in Parts; shows country instances of Dataset where each Tweet was written 1-9



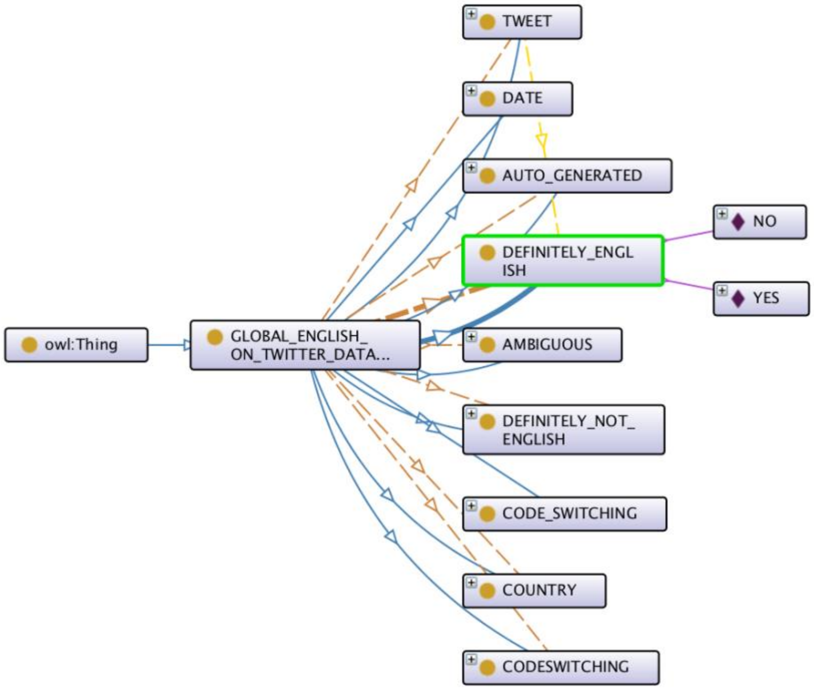
Ontology 4th Iteration shown in Parts; shows Yes and No instances in Def Not English Column of Dataset; these are encoded as 0 and 1 in Dataset and tells whether the tweet was written in a different language 1-10


THAI
FRENCH
HMONG
TELUGU
VIETNAMESE
SERBIAN
ARABIC
SLOVENIAN
SPANISH
LATVIAN
LANGUAGE
HINDI
GREEK
ENGLISH
PERSIAN
GUJRATI
PORTUGUESE
PUNJABI
CHINESE
URDU
ITALIAN

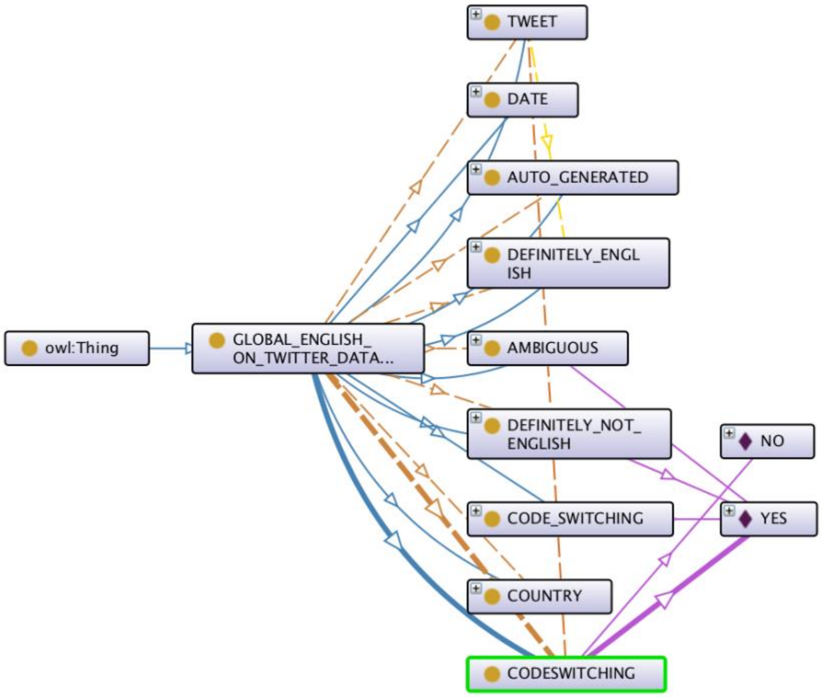
Ontology 4th Iteration shown in Parts; shows Language instances of Dataset; this was not a column in the dataset, but was created as part of the research 1-11



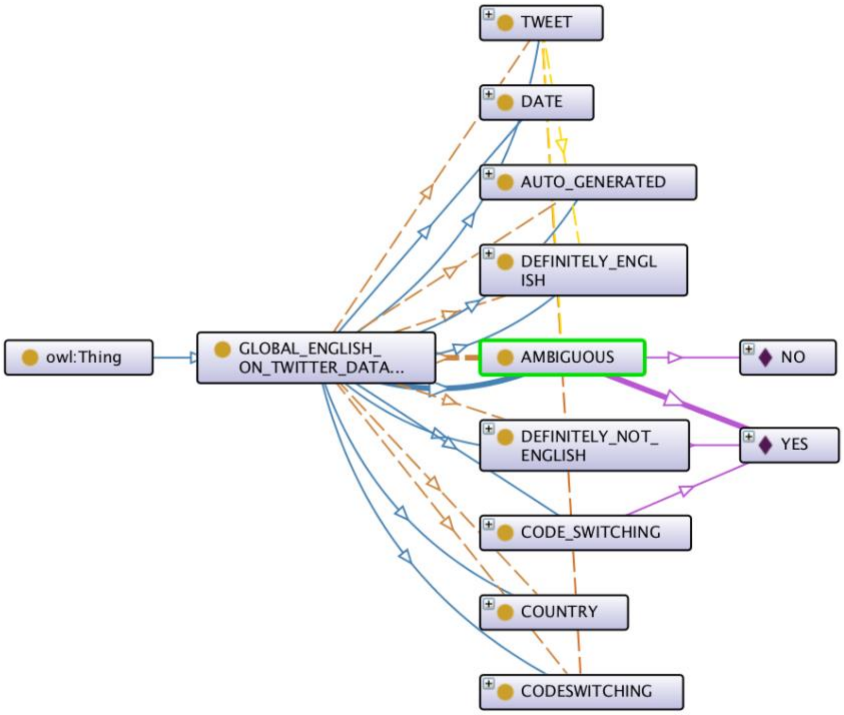
Ontology 4th Iteration shown in Parts; shows Yes and No instances in Autogenerated Column of Dataset; these are encoded as 0 and 1 in Dataset; shows whether the tweet was created by a computer 1-12



Ontology 4th Iteration shown in Parts; shows Yes and No instances in Def English Column of Dataset; these are encoded as 0 and 1 in Dataset; shows whether the tweet was written in English 1-13

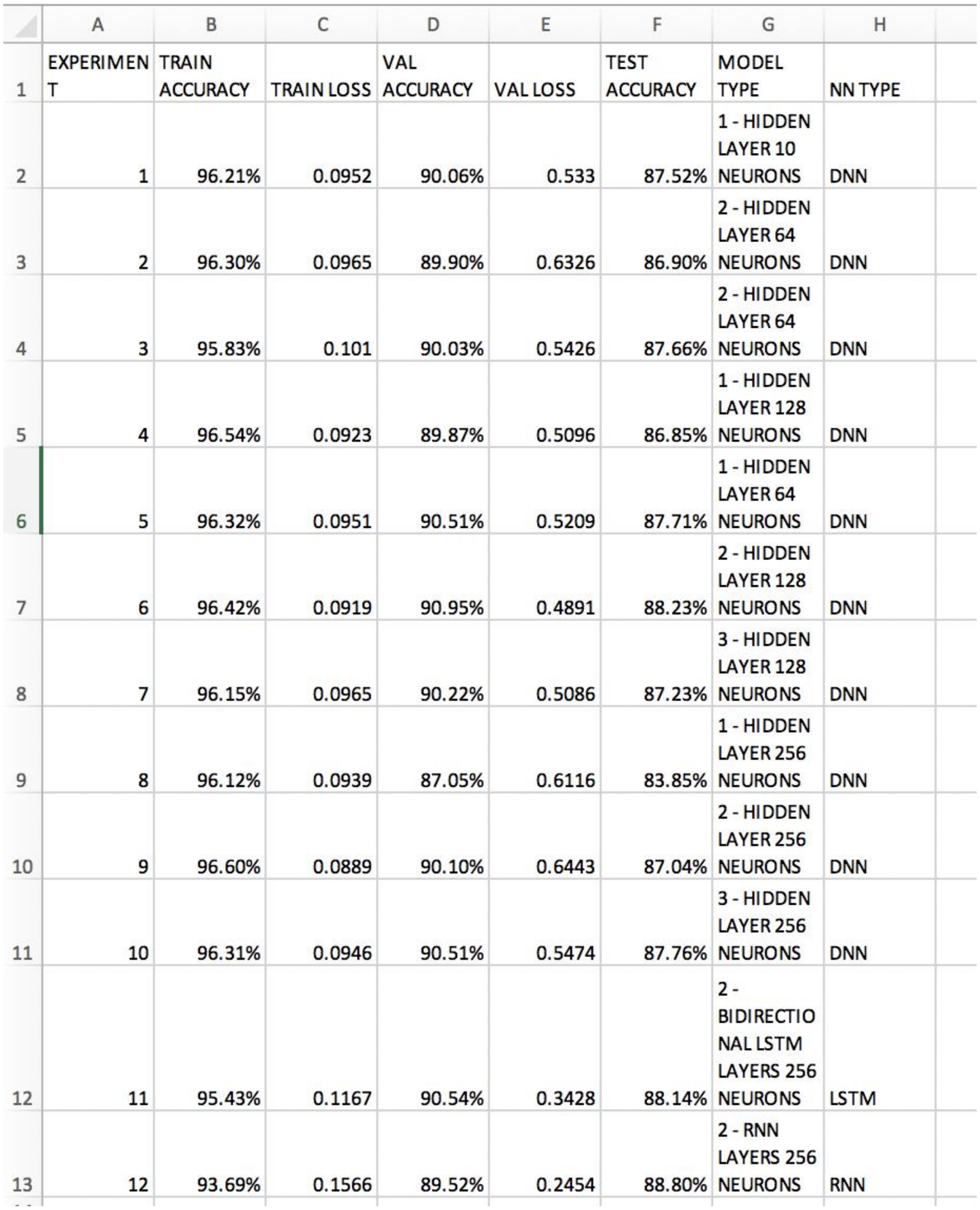


Ontology 4th Iteration shown in Parts; shows Yes and No instances in Code switching Column of Dataset; these are encoded as 0 and 1 in Dataset; shows whether a tweet contains more than one language 1-14



Ontology 3rd Iteration shown in Parts; shows Yes and No instances in Ambiguous Column of Dataset; these are encoded as 0 and 1 in Dataset; shows whether language is hard to determine

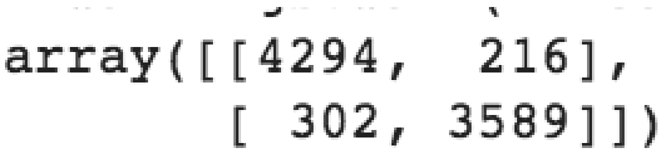
1-15



	A	B	C	D	E	F	G	H
1	EXPERIMENT	TRAIN ACCURACY	TRAIN LOSS	VAL ACCURACY	VAL LOSS	TEST ACCURACY	MODEL TYPE	NN TYPE
2	1	96.21%	0.0952	90.06%	0.533	87.52%	1 - HIDDEN LAYER 10 NEURONS	DNN
3	2	96.30%	0.0965	89.90%	0.6326	86.90%	2 - HIDDEN LAYER 64 NEURONS	DNN
4	3	95.83%	0.101	90.03%	0.5426	87.66%	2 - HIDDEN LAYER 64 NEURONS	DNN
5	4	96.54%	0.0923	89.87%	0.5096	86.85%	1 - HIDDEN LAYER 128 NEURONS	DNN
6	5	96.32%	0.0951	90.51%	0.5209	87.71%	1 - HIDDEN LAYER 64 NEURONS	DNN
7	6	96.42%	0.0919	90.95%	0.4891	88.23%	2 - HIDDEN LAYER 128 NEURONS	DNN
8	7	96.15%	0.0965	90.22%	0.5086	87.23%	3 - HIDDEN LAYER 128 NEURONS	DNN
9	8	96.12%	0.0939	87.05%	0.6116	83.85%	1 - HIDDEN LAYER 256 NEURONS	DNN
10	9	96.60%	0.0889	90.10%	0.6443	87.04%	2 - HIDDEN LAYER 256 NEURONS	DNN
11	10	96.31%	0.0946	90.51%	0.5474	87.76%	3 - HIDDEN LAYER 256 NEURONS	DNN
12	11	95.43%	0.1167	90.54%	0.3428	88.14%	2 - BIDIRECTIONAL LSTM LAYERS 256 NEURONS	LSTM
13	12	93.69%	0.1566	89.52%	0.2454	88.80%	2 - RNN LAYERS 256 NEURONS	RNN

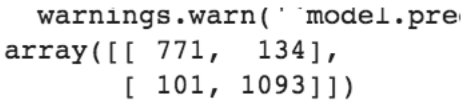
RESULTS OF 12 DEEP LEARNING EXPERIMENTS, DNN, RNN, LSTM NEURAL NETWORKS; SEE EXCEL SHEET FOR BETTER LOOK; RNN performed the best.

1-16



```
array([[4294,  216],
       [ 302, 3589]])
```

TRAIN SET CONFUSION MATRIX RNN MODEL 1-17



```
   warnings.warn(''model.pre
array([[ 771,  134],
       [ 101, 1093]])
```

TEST SET CONFUSION MATRIX RNN MODEL1-18

Analysis and Interpretation

A program Protégé was used for the 4th iteration of the ontology as it is a sophisticated way of making an ontology. In the fourth iteration the ontology was expanded to object properties and the relationships between the Tweets and Columns was revised.

In 1-7 through 1-15, the ontology is broken down into several screenshots for readability. The thinking behind this ontology was that there was a named dataset where the tweet was a parent class with the subclass of the column “Language”. The columns mainly “absolutely English”, “Not English”, “Code Switching” , “Country”, “Ambiguous”, and “Automatically Generated” had relationships with Tweet but were subclasses of Tweet and were thought of subclasses of the original dataset. There were columns which had two instances Yes and No encoded as ones and zeroes as explained above in the data section. The columns “country and language” was a list of languages and countries perceived to be in the dataset, so the list may leave off languages and countries.

The full dataset can be seen in 1-2, where tweets are represented by the “Tweet” column, and the other columns contain information about the tweet which can be thought of as metadata about the tweet. A tweet contained a one or more languages which can be shown by the “Code Switching” column. The other columns also contained information about the tweet such as whether the tweet was in “English or not”. The “Country” column was also important as it geotagged each tweet and with this information it would be easier to see which language the tweet is in. The "Ambiguous" column was used when it was not easy to decipher which language the tweet was in. Lastly the "Automatically Generated" column was used to describe whether the tweet was generated by a computer or person.

With this description of the dataset 1-2, it helped in drawing out the ontology as the tweet was the main component of ontology, so it was thought of as the parent class. Then there was subclass of the parent class which was the "Language" column, and all the other columns were thought of being object properties of the Tweet column, but they were not subclasses of the Tweet column because they were thought of as metadata as explained above. For example, a Tweet had a "country" but was not subclass as "country" was not a type of tweet, so these columns were considered a type of metadata of each Tweet. Each class contained instances. For example, the "Country" column contained instances such as United States, Mexico, India and the "Code-Switching" column contained Yes, and No Instances encoded as ones and zeroes. In the research a "Language" class was added but as shown in 1-2, there was no Language column. In order to move further with this research, it was important to add a language class to the ontology, as the tweets contained specific languages.

After constructing the Ontology, it was decided that the column to classify on would be the "definitely English" column. Since it was straightforward to use data preprocessing techniques such as removing punctuation and stop words in English, it was decided to classify on this column. The goal involved 12 experiments which could classify if the Tweet was in English or Not. The Results are shown in 1-16, where DNN, RNN, and LSTM were all tried. What was found is that the RNN Model performed the best among the 12 experiments with a very high-test accuracy of 88.80 percent. The Confusion Matrices were shown for both the Train and Test Sets for this model as seen in 1-17, 1-18, where it showed most of the data points being classified correctly.

Conclusions

In this research Binary Classification was conducted to decide whether Tweets were in English. In the research, a dataset of tweets was used in different language. The dataset consisted of 10,502 tweets. An Ontology was first conducted to recognize which columns could be used for classification. It was decided based on this ontology that the "Definitely English" column was the most straightforward to use for this research. The research went further in experimenting with deep learning models particularly RNN, DNN, LSTM. The best model was the RNN model with a test accuracy of 88.80 percent.

Directions for future work

In further research, it would be interesting to do further processing in other languages, and to explore different Python Packages for more language processing techniques. Spacy was useful as shown in 1-3 in the beginning as it gave a good idea of the languages in the dataset, since the dataset did not have the language labels. In further research, it would also be interesting to create the column for the language labels that Spacy spit out in 1-3, and use different preprocessing techniques, that would help classify the languages. Different packages would be needed for a specific languages for the preprocessing steps, which makes identification of languages a future research problem as this process would take time.

References

Asklov, E. (2016, November 8). *Code-Switching: The Weird And Wonderful Side Of Bilingual Communication*. Babbel Magazine. https://www.babbel.com/en/magazine/estoy-code-switching-like-loco-weird-and-wonderful-side-of-bilingualism#:%7E:text=Code%2Dswitching%20can%20also%20be,also%20switch%20to%20fit%20in.

Grossfeld, B. (2021, April 7). *Deep learning vs machine learning: a simple way to learn the difference*. Zendesk. https://www.zendesk.com/blog/machine-learning-and-deep-learning/#:%7E:text=The%20difference%20between%20deep%20learning%20and%20machine%20learning,-In%20practical%20terms&text=While%20basic%20machine%20learning%20models,they%20still%20need%20some%20guidance.&text=A%20deep%20learning%20model%20is,it%20has%20its%20own%20brain.

Mittal, Aditi. 2019. “Understanding RNN and LSTM.” Aditi Mittal. https://aditi-mittal.medium.com/understanding-rnn-and-lstm-f7cdf6dfc14e.

Noy, N. F., & McGuinnes, D. L. (n.d.). *What is an ontology and why we need it*. Protege. Retrieved May 16, 2021, from https://protege.stanford.edu/publications/ontology_development/ontology101-noy-mcguinness.html

O’Sullivan, C. (2020, October 21). *Deep Neural Network Language Identification - Towards Data Science*. Medium. https://towardsdatascience.com/deep-neural-network-language-identification-ae1c158f6a7d

SPRH LABS. 2019. “Understanding Deep Learning: DNN, RNN, LSTM, CNN and R-CNN.” SPRH Labs. https://medium.com/@sprhlabs/understanding-deep-learning-dnn-rnn-lstm-cnn-and-r-cnn-6602ed94dbff.

Tatman, R. (2017, September 26). *The UMass Global English on Twitter Dataset*. Kaggle. https://www.kaggle.com/rtatman/the-umass-global-english-on-twitter-dataset