

Final Report

“Identifying narrative text on reddit”

Cardiff University School of Computer Science and
Informatics

CM3203 – One semester individual project – 40
Credits

Student: Adriano Sole

Supervisor: Steven Schockaert

Abstract

“Identifying narrative text on reddit”

Neural network models for natural language processing are typically pre-trained on large text collections. This allows these models to learn world knowledge without the need for an explicit supervision signal. However, the kind of knowledge they can learn in this way crucially depends on the type of text collection that is used. For instance, Wikipedia is a common choice. By pre-training models on encyclopaedic text, they can acquire a lot of factual knowledge about the world. On the other hand, training models on narrative text (e.g. books or movie scripts) can be a better choice if learning common sense knowledge is the main goal.

In recent years, the social media site Reddit is increasingly being used for training natural language processing models. However Reddit contains documents covering a broad range of genres, including expository, narrative, and argumentative text.

This project explores the use of Reddit posts as training material for a number of different text classifiers.

I have gathered a selection of Reddit posts, 1336 posts from the Cardiff subreddit and 998 posts from random subreddits and labelled them all as either narrative or non-narrative. Using these posts I then trained each classifier in order to determine which one is best able to classify narrative text.

The main findings of this project are that the RoBERTa language model performs the best out of all the tested models. And the RoBERTa model would be the best choice to pursue further studies in filtering narrative posts in a Reddit corpus.

This project found that Reddit can be a suitable source of narrative posts, especially when the topic that the data is gathered from is restricted to one that contains realistic narrative posts.

Table of contents

Table of contents	3
Table of figures	4
Acknowledgements.....	5
Chapter 1: Introduction	6
1.1 What is a narrative.....	6
1.2 Using Reddit as a source of narrative text	7
1.3 Research questions	7
Chapter 2: Background	9
2.1 Gathering Reddit data.....	9
2.2 Natural language processing.....	9
2.2.1 Text classifiers.....	9
2.2.2 Dataset splitting	11
2.2.3 Hyperparameter optimization	12
2.2.4 Fitting text classifiers	12
2.2.5 F1 Scores	13
2.3 Python libraries	13
Chapter 3: Approach	14
3.1 Data collection	14
3.2 Labelling of collected data	14
3.3 Narrative examples	14
3.4 Implementation of code	17
3.5 Implementing tests	17
3.6 Implementing the RoBERTa method.....	19
Chapter 4: Results and Evaluation	20
4.1 Results.....	20
4.2 Evaluation	22
Chapter 5: Future work.....	36
Chapter 6: Conclusions	37
Chapter 7: Reflection on learning	38
References	39

Table of figures

Figure 1: Diagram of k-fold cross validation	12
Figure 2: An example of different fitting scenarios	12
Figure 3: Code screenshot of hyperparameter optimization	18
Figure 4: Graph of results for training each classifier on the Cardiff reddit data and testing it on the Cardiff reddit data.....	20
Figure 5: Graph of results for testing the Cardiff trained classifier on random subreddit data.....	20
Figure 6: Graph of results for training the classifiers on random subreddit and testing on random subreddit data.....	21
Figure 7: Graph of results for testing the random subreddit trained classifier on Cardiff data.....	21
Figure 8: Graph of results for testing the Cardiff trained RoBERTa model.....	22
Figure 9: Graph of results for testing the random subreddit trained RoBERTa model	22
Figure 10: Graph of comparison of F1 scores from using the Cardiff subreddit testing dataset with both types of classifiers	23
Figure 11: Graph of comparison of F1 scores from using the Random subreddit testing dataset with both types of classifiers	23
Figure 12: Image of text separation from post 1 in the false negatives section of table 3	31
Figure 13: Image of test text used to compute decision scores from post 3 in the false negatives section of table 3	32
Figure 14: Image of test text used to compute decision scores from post 3 in the false negatives section of table 3	32
Figure 15: Image of test text from post 3 in table 4 showing the addition of a few words	35
Table 1: Table of 5 False positive and 5 false negative results using the Cardiff LinearSVC classifier tested on Cardiff data	25
Table 2: Table of 5 false positive and 5 false negative results using the Cardiff LinearSVC classifier tested on random subreddit data.....	27
Table 3: Table of 5 false positive and 5 false negative results using the random subreddit LinearSVC classifier tested on random subreddit data.....	29
Table 4: Table of 5 false positive and 5 false negative results using the random subreddit LinearSVC classifier tested on Cardiff data	32

Acknowledgements

Thank you to my supervisor, Steven Schockaert, for assisting me throughout this project and providing valuable suggestions and criticism at all stages of its development.

Chapter 1: Introduction

This project is centred around the exploration of using Reddit posts as training data for natural language processing text classifiers and language models. These models are tested on two different datasets gathered from Reddit, a social media site, firstly to explore whether Reddit can be a source of narrative text for training language models, and secondly to analyse whether a language model can be trained to identify whether a Reddit post is a narrative.

The primary goal of this project is to determine how well a text classifier can determine if a particular post is a narrative or not, and directly linked into this is to examine what particularly does each text classifier determine to make a text a narrative.

Alongside this, the project requires a corpus of labelled reddit posts to be used for training the text classifiers, which means another goal of this project is to develop a method of determining whether a particular post is a narrative or not, alongside gathering the data itself.

The following lists some important outcomes of this project:

1. A method for labelling a piece of text as a narrative or not, including possible methods that may provide score improvements for text classifiers
2. A comparison of 5 different text classification models and their effectiveness at labelling a reddit post as narrative
3. An in-depth error analysis of the LinearSVC classifier and suggestions for avoiding possible errors in other text classification projects
4. A corpus of over 2000 Reddit posts labelled as narrative or non-narrative

To ensure the project was able to be completed within the semester the scope was limited to the above goals. While the scope could be expanded further by increasing the amount of data gathered which will improve on the amount of training data for the text classifiers.

1.1 What is a narrative

Narratives are pervasive throughout human interactions and text, existing in many forms from biographies to short stories. Generally, narratives will contain a retelling of something that has happened in the past that can be as minor as a personal experience or as large as a war, as Frederick Crews describes in 'The Random House Handbook' (Crews, 1977).

As this project is intrinsically linked with narratives, it is important to understand what a narrative actually is, as well as provide an example of what a simple narrative looks like since this project required the labelling of a large number of Reddit posts.

Narrative text is important in this project as it is an integral part of human conversation, it allows for language models to learn and understand about aspects of human communication that would be much harder to teach if a different type of dataset were used.

As mentioned in previously a narrative is a piece of text that will contain a retelling of something that has happened in the past, which can be as minor as a personal experience or as large as a war (Crews, 1977). This definition was used to make decisions about whether to label a certain post as narrative or not.

A simple example of a narrative post is the following.

“Went through there today and there's quite a few lorries outside, camper vans inside and lots of stuff set up in there with people wearing professional gear!”.

And an example of a non-narrative post is the following.

"Hi all! I'm looking at moving to Cathays next academic year with some other students and I was wondering what wifi is the best for the area? We'd only need a 12 month contract and we don't want terrible Wifi because we're all studying online at the moment Thank you!!"

A previous study by Andrew Gordon and Reid Swanson (Gordon & Swanson, 2009) delved into the labelling of weblog posts as stories or not, and the data labelling approach they used involved having a team of annotators that would go through a set of posts to label them individually and then compare the results of each rating.

Unfortunately, as this project was not as large in scope as this study, the labelling was done entirely by me. This meant that the likelihood of human error was much higher, and this may have affected some of the results that I had gathered during the project.

1.2 Using Reddit as a source of narrative text

This project uses data gathered from Reddit, a social media forum with topics ranging across many boards known as subreddits, a subreddit being a subsection of the forum which focuses on one particular topic such as /r/Sports, /r/Programming, or /r/Gaming. In particular, for the scope of this project I have focused on data from /r/Cardiff, as well as a random selection across all subreddits. In total I have gathered roughly 2000 posts, 1000 from the Cardiff subreddit and 1000 from random subreddits.

The data gathered from Reddit was used to train a variety of text classifiers. Reddit is a good source for narrative text as there are many types of subreddits that will contain people who are trying to convey a story or past event to promote discussions and look for opinions. In particular the Cardiff subreddit will likely contain narrative posts as it is expected that people will be writing posts about an event that has occurred in their daily lives around town.

I chose to gather data from both the Cardiff subreddit and a random set of subreddits as this would provide some comparison on which subreddit would have the better score in identifying narrative posts.

There are a number of challenges involved when identifying narratives, in particular the issue of deciding how to label each piece of data. As far as narratives can be distilled down to basic parts there is still an issue involved in deciding how to label posts that contain both narrative and non-narrative aspects, solutions to this issue are discussed further in the paper.

1.3 Research questions

This project will attempt to answer a number of questions relating to the text classifiers mentioned above.

1. It will be important to determine which classifier is able to achieve the highest performance, and with which dataset it does this

2. Analyse if using different datasets for training and testing will provide better results for the classifiers
3. This project will also attempt to analyse and compare why certain classifiers perform better than others in this scenario
4. Examine what particular things that a classifier is able to determine about narratives based on the given datasets

Chapter 2: Background

2.1 Gathering Reddit data

The project required that I gathered a reasonably large number of Reddit posts which would be too time consuming to gather by hand, therefore I decided to use one of a number of tools to download posts onto my local machine.

The Pushshift API is a tool designed by moderators on the `/r/datasets` subreddit which provides search functionality for Reddit posts, including the ability to specify what subreddit you want to filter by and the contents of a post.

Using this tool I was able to gather the textposts for two different sets of data: one set contains posts gathered from the Cardiff subreddit, and the other contains posts taken at random from all of Reddit.

While searching for tools to download Reddit posts I came across a number of alternatives to directly requesting the Pushshift API.

The tools that I found were the PSAW API and the PRAW API, both of which allow for the downloading of Reddit posts. In particular I chose not to use these alternatives as the Pushshift API was much easier to use.

In particular the PRAW API would allow me to have a much more control over which posts to obtain by filtering them by certain values such as the amount of upvotes a post has, however I still chose to use the Pushshift API as it has the basic functionality that covers exactly what I needed for this project.

2.2 Natural language processing

2.2.1 Text classifiers

There are many different text classifiers that are commonly used in the field of natural language processing. I have chosen the following 5 to be tested in this project due to their ability to provide a spread of results using the Reddit dataset.

Each of these classifiers have performed differently, in some cases better or worse than others. Reasons behind why these classifiers may perform in the way they have will be explained in the following paragraphs.

In order to test the following classifiers I had to convert the dataset that I had gathered into a vector representation. To do this I used two different methods, the first method was to convert each post into a word vector using a Python library called Spacy

The second method was to use the built-in methods in the Scikit-learn Python library to convert posts into a count vector representation, which creates a matrix of the number of features in the text. This representation is not as useful for this project as it only compares the number of features in one post to another, so what I instead did was to use a tf-idf transformer to turn the count vectors into a tf-idf representation.

Tf-idf stands for term-frequency times inverse document-frequency, this is a method designed to scale down the impact of repeated features in a post while making sure that features which occur less often are not assigned less value.

The Support Vector Machine (SVM) classifier is a classifier that takes vectors as input, in this case a word vector representation of the dataset and maps them to points in space. It then compares the testing examples with the mapping it has created and predicts which category they belong to.

SVM classifiers can provide effective classification when used in high dimensional spaces, and in this project the Linear Support Vector Classifier (LinearSVC) is used as it is much more robust and faster than the SVM classifier. This means that it is less prone to overfitting data, a process where the classifier is able to understand the training examples very well but fails to adapt this to testing examples.

The K-Nearest Neighbour classifier (kNN) is a classifier which stores word vector representations of training data. This classifier makes predictions by taking a post and examining the k nearest examples to determine which ones have the most similar vectors, with this information it assigns a post a certain label based on the highest number of similar examples. Despite being a simple algorithm, the nearest neighbour's algorithm has been successful in a large number of classification and regression problems. (Goldberger, et al., 2005)

This project will also be using the Multinomial and Bernoulli Naïve Bayes classifiers. A Naïve Bayes classifier is based on applying Bayes' theorem, the way in which the Naïve Bayes classifier works is that it assumes a document belonging to a certain label will contain words in a specific proportion, using this information it then estimates the probability of a word appearing given that it belongs to a certain label. The naïve aspect of the Naïve Bayes classifier is that there is an assumption that the probability of seeing a given word is independent of other words appearing.

The Multinomial Naïve Bayes classifier implements the naïve Bayes algorithm for multinomially distributed data, while the Bernoulli Naïve Bayes classifier implements the naïve Bayes algorithms for data that is distributed according to multivariate Bernoulli distributions. The main difference between these two algorithms is that the Bernoulli naïve Bayes classifier explicitly penalizes the non-occurrence of a feature that is an indicator for a particular label.

Both of these versions of the Naïve Bayes classifier have been selected in order to compare which one provides the best performance since the Reddit dataset will likely contain varying lengths of individual posts. This is important to compare as the Bernoulli classifier tends to perform better at smaller text sizes while the opposite is true for the Multinomial classifier (McCallum & Nigam, 1998) and this is because the Bernoulli classifier will penalise a post more if it is longer, due to the chances of a feature appearing that is not related to the label of that post.

Of these four classifiers, the LinearSVC classifier is expected to perform the best. This is because the LinearSVC classifier tries to classify text based on something called a hyperplane, a linear separation drawn between all of the examples in the dataset that places positive examples on one side and negative examples on the other. The classifier then tries to determine what to classify testing dataset by comparing it to the positive and negative examples in relation to the hyperplane.

This means that the LinearSVC classifier is able to make assumptions about what places a post in the positive or negative plane, meaning that it will try to understand what aspects of a post will make it more likely to be a narrative vs a non-narrative.

In comparison classifiers like the kNN classifier create a very specific interpretation of the training dataset that does not have any constraints, meaning that it can create wildly different interpretations of the dataset each test. What this also means is that the classifier will be able to

remember specific posts and their labels but may struggle to understand what exactly is making a post be labelled a certain way which gives it a high accuracy score but means that it will struggle to label new unseen examples.

Another language model that I will be testing alongside the four classifiers above is RoBERTa.

RoBERTa is an optimization of the Google-created BERT language model (Facebook AI, 2019). RoBERTa is a deep neural network that has been pre-trained on a very large text collection in order to be able to predict masked words. In doing this RoBERTa is able to achieve a degree of language understanding before it has even seen the training set that I will fine-tune it on.

RoBERTa itself is pretrained using Masked language modelling. This means that the model randomly masks 15% of the words in the input and runs the masked sentence through the model in order to have it predict the masked words. Using this the model is able to learn an inner representation of the English language.

The outputs of the model are used to train a linear classifier, on top of all the layers of fine-tuning that the neural network is doing. This means that while the linear classifier is being trained the layers of the neural network is having the weights fine-tuned to ensure that the model focuses more on the language used in the Reddit datasets.

The RoBERTa model was chosen as it was expected to provide a much better score than the other classifiers due to the way the model works. Since the RoBERTa model is pre-trained on large amounts of data it already has a baseline understanding of human language and text, which means that the classifier will be able to have knowledge of certain aspects of writing before it even sees the training data. This gives RoBERTa a distinct advantage over the other classifiers that I have selected and is why it is expected to perform the best out of them all.

2.2.2 Dataset splitting

An important aspect of using the above classifiers with a dataset is to ensure that the data has been split correctly into three different sets, the training set, the validation set, and the testing set.

These three sets ensure that the classifiers can be trained and tested without causing issues with cheating, for example using the same dataset for both validation and testing is considered cheating, as you are using your final test multiple times to determine what the best parameters for that particular dataset are when it should be more generally decided on a validation set.

The data is split into different sets in various ratios, for example 72% training data, 8% validation data, and 20% testing data. This allows for the majority of the data to be used for training the classifiers and provides a balance between validating the classifier model and testing the final model.

A possible method to use with splitting data is to use cross validation, a system where instead of selecting a flat percentage for each set you would instead rotate which percentage of the dataset is used for each set.

For example in a 4-fold cross validation the dataset is split into four equally sized sections, in the first iteration of training and testing the first part of the dataset would be used as the testing set while the last 3 parts would be used as the training set. After gathering results for that iteration you then rotate which parts are used for the training and testing sets, i.e. the second fourth of the dataset would be used as the testing set while the other 3 parts are used as the training sets. This example is illustrated in the figure below.

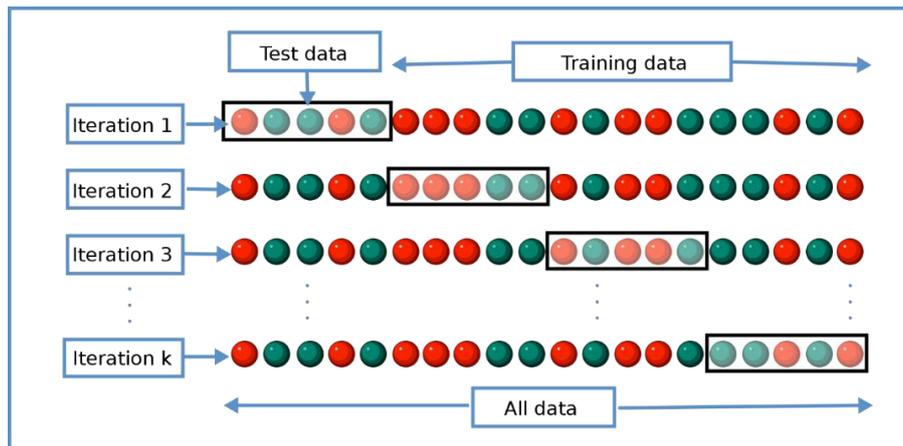


Figure 1: Diagram of k-fold cross validation

There are two main reasons to use k-fold cross validation, the first being that this method uses the entire dataset for testing. Since the method will iterate over the whole dataset it means that the conclusions that you can make will be more reliable as you are using more examples for testing. Another reason to use k-fold cross validation is that the model is trained k times and the average performance is taken from those k iterations, this reduces the chance of random variations in performance in cases where the classifier is particularly lucky or unlucky.

2.2.3 Hyperparameter optimization

Most text classifiers rely on something called hyperparameters, which are customisable variables that will affect the way that the classifier processes data. Many classifiers will have default hyperparameter settings, however in order to achieve the best results from each classifier it is important that I tune the hyperparameters using data that is different from the testing data, since it would be considered cheating to base the parameters off of data that I have already tested.

2.2.4 Fitting text classifiers

The process of fitting or training a classifier is the process where the classifier is given a set of training data and the data is fitted to the classifier, after which it can then be used to make predictions on testing data.

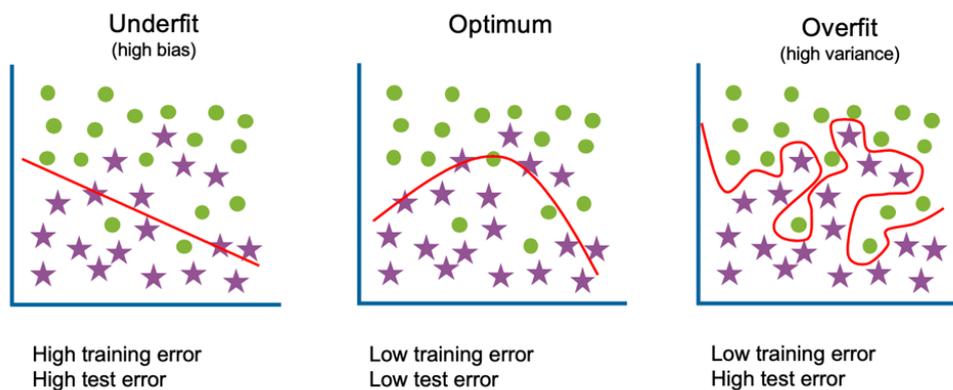


Figure 2: An example of different fitting scenarios

Obtaining a good fit for the training data is a key part of this project, as it will impact the scores that are obtained at the end. In order to improve the fitting of the classifiers I have optimized the hyperparameters to obtain the best possible F1 score.

2.2.5 F1 Scores

A common way of determining performance in natural language processing tasks is to calculate the F1 score of the predictions that have been made. This is done by using the following formula:

$$F_1 = \frac{2}{\text{recall}^{-1} + \text{precision}^{-1}} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} = \frac{\text{tp}}{\text{tp} + \frac{1}{2}(\text{fp} + \text{fn})}$$

Where tp stands for true positives, fp stands for false positives and fn stands for false negatives.

The F1 score is a good method of determining the performance of a language model as it combines the precision and recall scores of a model to determine its overall score. This means that the F1 score determines the ability for a model to select relevant examples whilst minimising the number of incorrect examples it selects. I have used the F1 score metric in this project as a baseline for determining which language model is the best.

2.3 Python libraries

This project will use a number of different python libraries in order to test the Reddit data.

The main library that I will be using is Scikit-learn, a huge open source Python library that contains a multitude of different machine-learning tools that can be used. (Pedregosa, et al., 2011)

In particular, this tool will be used to split the training data, create the classifier models that I will be testing, and calculate the F1 scores of each classifier.

Simple Transformers is another library that I will be using, particularly for its simplicity and ability to allow me to test the RoBERTa classifier in my code.

The Simple Transformers library is based on the Transformers library. Simple Transformers is designed to simplify the process of using the Transformers library to make training and testing the machine learning models much easier for simpler tasks. (Rajapakse, 2020)

Since this project does not have a huge scope, the Simple Transformers library will allow me to save a large amount of time by simplifying the whole process of using the RoBERTa classifier.

Spacy is an open source Python library that normally allows for performing NLP tasks. It is considered to be an industry standard and is specifically designed for production use. (Spacy, 2020)

For my project specifically Spacy has been used to convert my dataset into word vectors, a type of text representation that is used to encode the meaning and context of a sentence into a vector format so that a classifier can read it. This representation allows for the classifiers to understand if a particular sentence is closer in the vector space to another, which will influence how it labels posts (Jurafsky, et al., 2020).

Chapter 3: Approach

3.1 Data collection

Collecting the Reddit data was an integral first step in fulfilling the research goals of the project and was something that was a little more difficult than I initially expected it to be.

Using the Pushshift API I was able to successfully download Reddit posts. Initially my goal was to download 1000 posts from the Cardiff subreddit and 1000 posts from random subreddits. However there was an issue with downloading a certain number of posts per request of the API. The tool seemed to only let me download a random number of posts whenever I sent a request which would vary between 100 and 150 posts.

In order to work around this issue, I chose to create a for loop to iterate over a period of 2 years, capturing roughly 100 posts per month for that period. The end result of this strategy gave me 1336 posts for the /r/Cardiff dataset and 998 posts for the random subreddit dataset.

In addition, I performed some minor text processing on both datasets to remove special characters as they were causing issues with the training of the classifiers, and I also setup the data in a json format of 'post' and 'label' for each entry so that it would be in the correct format to be added to a dataframe object in my Python code.

3.2 Labelling of collected data

The next step I took in the project was labelling both datasets that I had collected.

The approach I took to labelling the data that I gathered is similar to the one used in a previous study done on weblog stories (Gordon & Swanson, 2009), in which their data was manually labelled and compared between a team of labellers.

Considering the scale of this project, I chose not to gather a team of people to label the Reddit data, as such I labelled it all by myself. A possible drawback of choosing to label the data this way is that it meant that I may have mislabelled data as I did not have anyone to compare my dataset with.

I read through each individual post in my captured Reddit data and observed whether they could fulfil the following criteria:

- Is the primary focus of the post a retelling of an event that has occurred?
- Does the post contain a clear chronological order of events that occurred?

Using these criteria, I found there to be a total of 241 narrative posts and 1095 non-narrative posts in the Cardiff dataset, while the random subreddit dataset contained 235 narrative posts and 763 non-narrative posts.

This meant that I had a fairly imbalanced dataset, with there being roughly one narrative post per 5 non-narrative posts. This may have caused issues in my project, as having fewer narrative posts meant that there were less examples for the language models to train on.

3.3 Narrative examples

The following posts provide some examples of where I would label a post as narrative vs non-narrative.

"I did my first one last week at Bute Park and again today. I am a total beginner. There were so many people there! It's absolutely brilliant and totally free. The marshalls all give up their time too. I Might try the Grangemoor one next week as Bute Park is so busy. Anyone else go Or New to It like me?"

This post was labelled as a narrative, the overall goal of this post is for the person writing it to narrate what happened during an event at Bute park in the previous week. This resulted in a post which is mostly in the past tense, which describes an event and gives us the timeframe for when that event occurred.

"Hi all! I'm looking at moving to Cathays next academic year with some other students and I was wondering what wifi it the best for the area? We'd only need a 12 moth contract and we don't want terrible Wifi because we're all studying online at the moment Thank you!!"

This post was labelled as non-narrative as it doesn't contain any recollection of a past event or any chronological order for events to occur.

"I cycled down Senghennydd Rd for the first time since the bike lane was put in and it looks like a perfect storm for a bike pedestrian collision. Straight bike lane with a downward gradient you can easily get up 20 or 30 miles an hour on. And A never ending supply of students staring at their phones who think of the bike lane as an extension of the pavement. It just a matter of time till it happens."

This post contains the retelling of a person cycling down a road, while the second half of the post could be considered non-narrative the overall sentiment of the post is to describe the persons feelings at the time of the event.

"So we have been having a bit of a clear out, what with the latest lockdown, but have realised that after bagging a load of stuff, we have nowhere local to take it. Can anybody let me know if there are any charity donation places that are still open?"

This post contains a mixture of narrative and non-narrative, with the first sentence containing a series of events about a person clearing out some items and needing to take it somewhere. While the second sentence contains a question relating to the narrative that they have just described. This overall pushed me towards labelling this post as a narrative.

"Seems like the coronavirus is starting to ramp up around the world. At what point will you start preparing? I pray the people of Cardiff wont panic if there is an outbreak here. But given recent events in Italy, it's likely panic would spread quickly. I've ordered some face masks a few weeks back and know a few people who have done the same. Call it silly but worse case scenario it's better to be prepared than not. It's time to really start thinking about what you would do. Peace"

This shows another example of a post with a mix of narrative and non-narrative text. The post begins with mostly a discussion around the coronavirus, with the post leaning towards a narrative in the 6th sentence. Overall this post leans more towards non-narrative than it does to narrative and so I labelled it as non-narrative.

"Does anyone know of or recommend a book club in the Roath area? If there isn't one, would anyone be interested?"

This post was labelled as non-narrative as it clearly does not contain any recollection of a past event. It is entirely a person asking questions.

"I'm a student but don't have a dentist here or at home - can anyone recommend a good one in Cardiff, preferably nearer to Cathays or the centre but can travel. Doesn't have to be NHS"

This post similarly does not discuss a past event and is instead about a student who is looking for dentist recommendations.

"So recently I have been playing against a lot of no-name hackers losing 4-0 or 4-1, etc. All these hackers show up as no name if you click on their profile or on the leaderboard or kill feed. Learned if you friend the hacker it says you sent a friend request to said hackers actual username. I was wondering why so many recently. I play at the silver-gold range and had hackers maybe 5/7 recent games."

This post is an example of a narrative post from the random subreddit dataset. The information present in this post is very specific, coming from a subreddit about a game. This post was labelled as a narrative as it contains a retelling of an event where the player has been playing against hackers.

"Everything was working fine yesterday, when I tried booting it up today all i got was a white box that read, "General Error. (0xE0010160)" I tried scan/repairing but nothing works."

This is another example of a post where I have labelled it as a narrative. The overall point of the post is to tell story about how the poster is having errors with some hardware that is giving them errors, with a clear mention of the time frame that the story occurred in.

Many posts in the Cardiff subreddit contain questions from people seeking recommendations, such as restaurants or neighbourhoods to purchase houses in. While the random subreddit dataset contains many different topics across a range of subjects.

There were a number of issues in the labelling process regarding the choice of labelling posts as a binary narrative or non-narrative choice, since there were certain posts which could be considered to be a mix of both narrative and non-narrative. In these cases I opted to focus generally on the overall percentage of the sentences that could be considered narrative or non-narrative, in this way I was essentially using my personal judgement to give a post an overall rating.

This may have caused some issues in the results of the project with classifiers having issues discerning what to label the post as, considering a classifier would be unable to judge a post in the same way as a human.

To overcome this it would have been a better idea to label posts with a scale from narrative to non-narrative, which would allow for more lenience in the edge cases when labelling posts that covered both categories. This strategy could provide better results as it would enable classifiers to understand what parts of a post are more narrative than non-narrative.

There could be issues with this strategy as well, considering what parts of a post contribute more to it being a narrative vs non-narrative would end up in the same position as I was with having to make a personal choice over what score to give a certain post.

3.4 Implementation of code

After finishing labelling the datasets, I then went on to begin writing the code that I would be using to gather results on the classifiers that I tested.

Some sections of my code were sourced from the Scikit-learn tutorial for implementing text document classification using a number of different classifiers (Scikit-learn, n.d.).

I used Python for the project as it has many machine-learning libraries that come with many classifiers pre-installed, as well as important methods that would cut down on the time spent coding.

Initially the goal of this project began with testing 4 classifiers, the LinearSVC, KNearestNeighbour, Bernoulli naïve Bayes, and multinomial naïve Bayes. At a later date the RoBERTa training model was added as another method of machine learning to see if it provided better results than the other 4 classifiers.

Importing the data was a simple process of using a method from the Pandas library to read the json files into a dataframe object. This enabled me to avoid any issues with creating my own data structure for the posts.

Using the built in scikit-learn method for splitting data, `train_test_split`, I specified the following ratios of data for the method to separate. I chose to split the data into 72% training, 8% validation, and 20% testing.

After separating the data I then used Spacy to convert the data into word vectors, which I used to train the LinearSVC and KNeighbours classifiers. I sourced the code used to convert my data into word vectors from an online blog (Levengood, 2020).

For both naïve Bayes classifiers I instead used a built in count vectorizer in Scikit-learn that I passed the dataset to which converts the data into a count vector, which I then passed through to a tf-idf transformer method in Scikit-learn that converts the count vector into a tf-idf representation.

3.5 Implementing tests

With both formats of datasets ready I then moved onto the training process for each classifier.

The first step in the process of training the classifiers that I would use for the project was to optimise the hyperparameters for each one.

The language models that I used each have their own hyperparameters. The LinearSVC classifier uses a hyperparameter called C, the values that I tested for this were .01, .1, 1, 10, and 100. The KNeighbours classifier has a hyperparameter called n neighbours, the values that I tested for this were 3, 5, 10, 15, 20, 30, 44, and 70. The Bernoulli and Multinomial naïve Bayes classifiers have a

hyperparameter called alpha, the values that I tested for both of these were .000001, .00001, .0001, .001, .01, .1, 1, 10, and 100.

The above selection of hyperparameter values has been selected to provide a wide range of values for each possible classifier. The values are generally 10 times higher for each one, as the impact of smaller increments makes little difference compared to larger jumps in size. Using these values I was able to search for the best scoring parameter which provided the best performance for my datasets.

Originally I tried to use one of the Scikit-learn methods for hyperparameter optimization, however I realised after implementing the code that the Scikit-learn gridsearch method looks at the F1 score across all the labels in the dataset. This was an issue due to the project being solely focused on the scores for the narrative label, and therefore to overcome this I created my own code for optimizing the hyperparameters for each classifier.

```
41 Cs = [.01, .1, 1, 10, 100]
42 bestScore = 0
43 bestParam = ''
44 for singleC in Cs:
45     tuner = LinearSVC(max_iter=100000, C=singleC)
46     tuner.fit(cardiffData_train_x_wvec, cardiffData_train_y)
47     predicted = tuner.predict(cardiffData_tune_x_wvec)
48     f1Score = metrics.f1_score(cardiffData_tune_y, predicted, pos_label='narrative')
49
50     if f1Score > bestScore:
51         bestScore = f1Score
52         bestParam = singleC
```

Figure 3: Code screenshot of hyperparameter optimization

As can be seen from the above figure, I used a simple for loop to iterate over a list of possible hyperparameters, in this case the C value for the LinearSVC classifier, and trained the classifiers using each hyperparameter value.

I then used the classifier to predict results using the validation set and calculated the F1 score for that iteration, storing the highest found F1 score and parameter choice per iteration which would be used in training the actual classifier.

After finding the best hyperparameter for a certain classifier I then trained it using that parameter and used it to predict results using the testing dataset. With those results I then calculated the F1 score to see the results of training that classifier with either the Cardiff or random subreddit datasets, and then testing the data on either the Cardiff or random subreddit datasets.

The above process was done for four different tests: Once to train the classifier on Cardiff subreddit data and test it on Cardiff data; Then to test the Cardiff subreddit trained classifier on the random subreddit data; Once to train the classifier on the random subreddit data and test it on the random subreddit data; And then to test the random subreddit trained classifier on the Cardiff subreddit data.

These tests were conducted in order to determine whether the Cardiff dataset would provide a better score than the random subreddit dataset. The tests also provided insight on whether the use of a more specific topic would give a better overall score than a dataset that is more general.

Using this method I gathered data for all four classifiers using both datasets and was able to compare the results of each scenario to see in which case the classifiers performed the best.

3.6 Implementing the RoBERTa method

After considering the results that I had achieved by this point to be lacklustre, I then opted to try using the RoBERTa training model to see if I could achieve better results with a more powerful neural network.

Implementing the code for RoBERTa was a much more challenging endeavour than the previous section as it is typically used with a graphics card to help speed up the process of training the neural network model. The primary issue with this was that regardless of the dependencies that I installed the code would not function, therefore I had to perform the training aspect of the model using only my computer's CPU which was much slower than expected.

To utilize the RoBERTa model I installed a Python library called Simple Transformers. This library massively reduces the effort involved with using the RoBERTa model down to only a few lines of code.

The process for using RoBERTa was very similar to the process for the other text classifiers, I had to change some of the formatting for my datasets, but the process generally followed the same strategy.

I went through the process of optimizing the hyperparameters for RoBERTa, in this case I chose to optimise the learning rate with the following values $2e-5$, $3e-5$, $4e-5$, and $5e-5$. I was able to predict the data using the four different processes mentioned above, which then gave me results for the RoBERTa training model.

Chapter 4: Results and Evaluation

4.1 Results

The results I gathered for each classifier had a fairly large degree of variance across the different dataset tests.

The first set of results is for training each classifier on the Cardiff reddit data and testing it on the Cardiff reddit data.

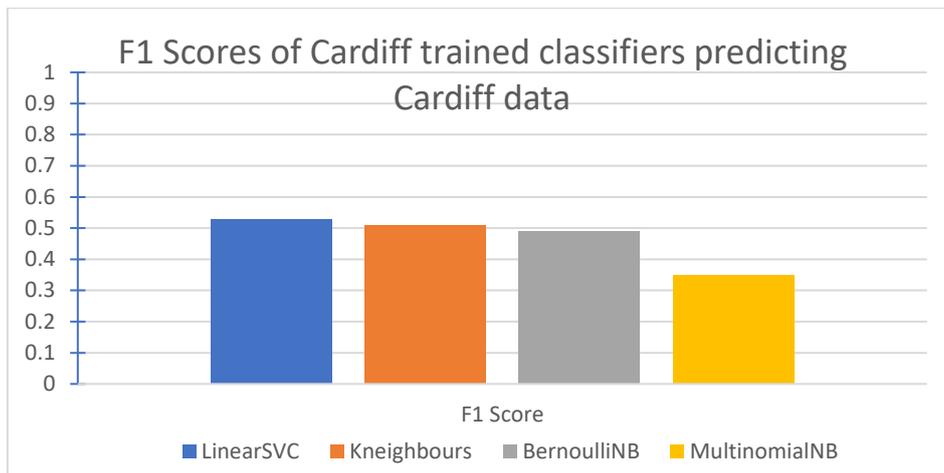


Figure 4: Graph of results for training each classifier on the Cardiff reddit data and testing it on the Cardiff reddit data

This data indicates that LinearSVC had the best F1 score with a value of 0.53, while the other classifiers had 0.51, 0.49 and 0.35 for KNeighbours, BernoulliNB and MultinomialNB, respectively.

The next set of results is from using the Cardiff trained classifier to predict random subreddit data.

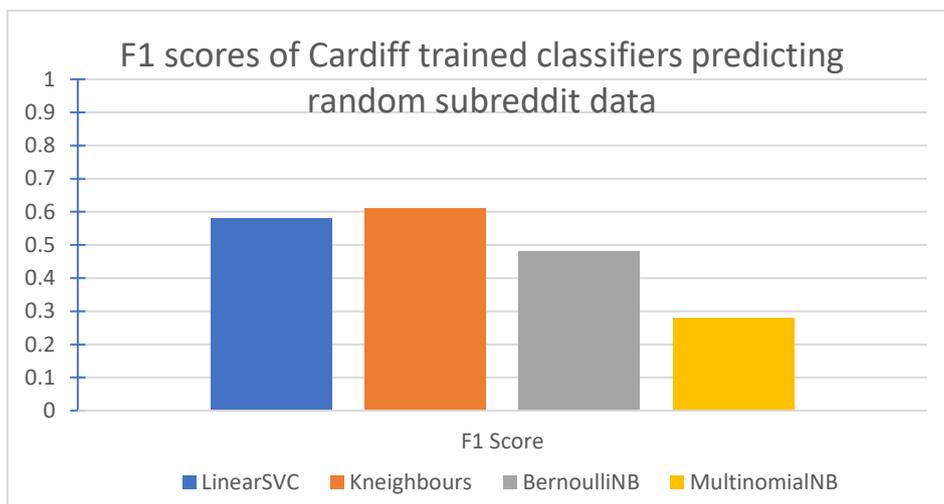


Figure 5: Graph of results for testing the Cardiff trained classifier on random subreddit data

These results show a notable increase in F1 scores for both the LinearSVC and KNeighbours classifiers, however there is a drop in score for both Naïve Bayes classifiers.

Results for the random subreddit classifier predicting random subreddit data show similar numbers to the results in figure 3.

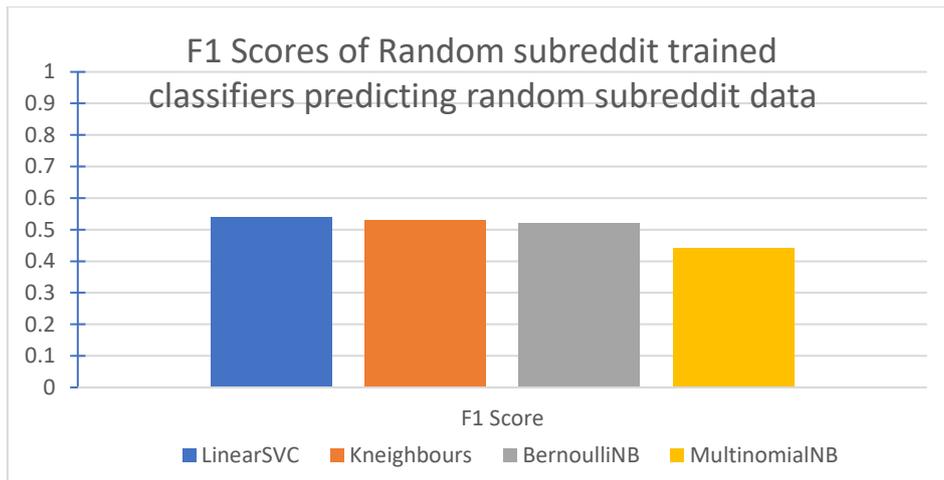


Figure 6: Graph of results for training the classifiers on random subreddit and testing on random subreddit data

One of the main differences compared to figure 3 is that the Multinomial Naïve Bayes classifier was able to score higher by 0.16.

The next set of results shows the effects of testing the random subreddit classifier using the Cardiff subreddit data.

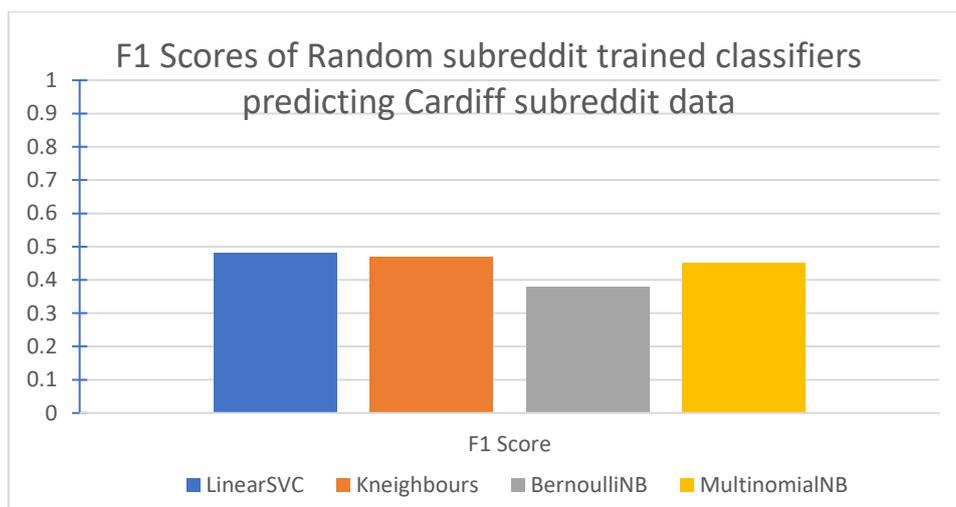


Figure 7: Graph of results for testing the random subreddit trained classifier on Cardiff data

These results are noticeably lower than the results shown in figure 4, with the highest score in this case being 0.48 for the LinearSVC classifier.

The next set of results are from the predictions gathered using the RoBERTa training model, with the same method of prediction used in the previous set of results. These are the results for the Cardiff trained RoBERTa model.

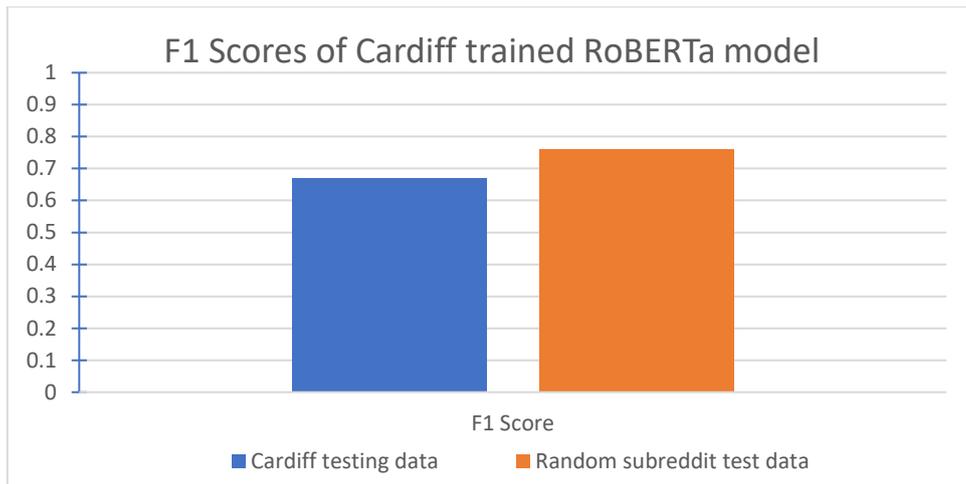


Figure 8: Graph of results for testing the Cardiff trained RoBERTa model

These results are much higher than the previous 4 classifiers, and what is interesting is the large difference in score between predicting Cardiff test data compared to predicting random subreddit test data.

The following graph shows the F1 scores for the random subreddit trained RoBERTa model.

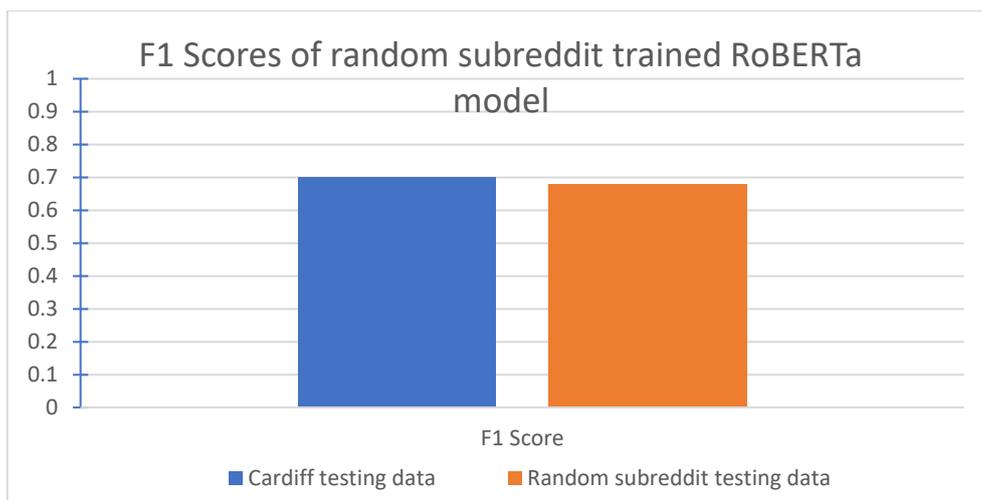


Figure 9: Graph of results for testing the random subreddit trained RoBERTa model

The data here is almost the inverse of the results shown in figure 7, with the Cardiff prediction having a higher F1 score for the random model compared to the random prediction for the Cardiff model.

4.2 Evaluation

One of the initial goals that this project had in mind was to compare and determine which language model would provide the best possible score, as it would provide interesting results of whether a more general training set like the random subreddit set would be better than a specific set like the Cardiff subreddit dataset.

This is important as it would provide insight on areas of Reddit that would make for a better source of narrative text, which could influence which subreddits are chosen in future studies of using Reddit for NLP tasks.

Overall, I have observed the outcome that for the tests where the classifiers were trying to predict Cardiff data, the best results came from the Cardiff trained LinearSVC classifier. And in general the best results for this test were from the Cardiff trained classifiers This is illustrated in the following figure.

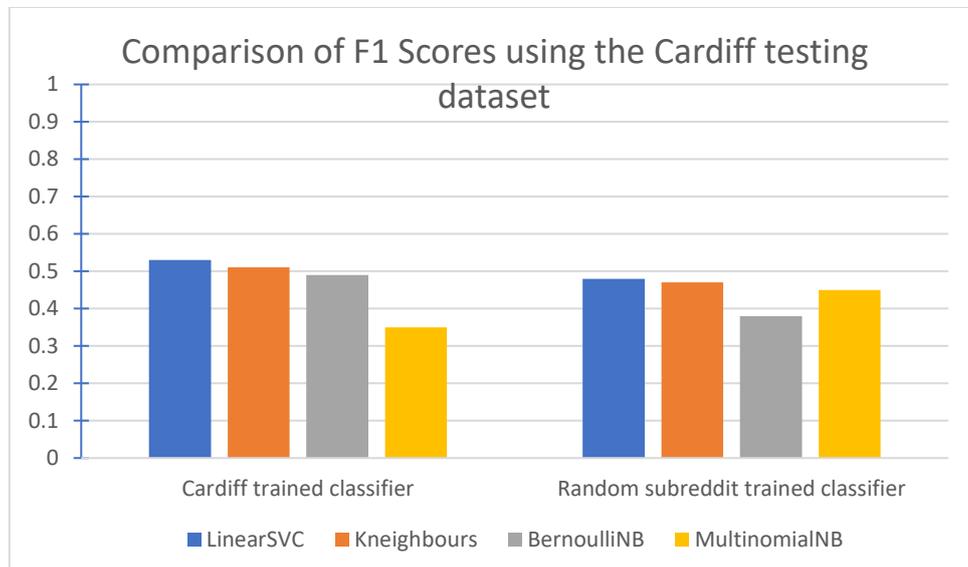


Figure 10: Graph of comparison of F1 scores from using the Cardiff subreddit testing dataset with both types of classifiers

In the test case where the classifiers were trying to predict the random subreddit dataset, the best results came from the Cardiff trained KNeighbours classifier. However overall the random subreddit classifiers achieved a higher average score of 0.51 vs the average of 0.49 from the Cardiff trained classifiers. As can be seen from the following figure.

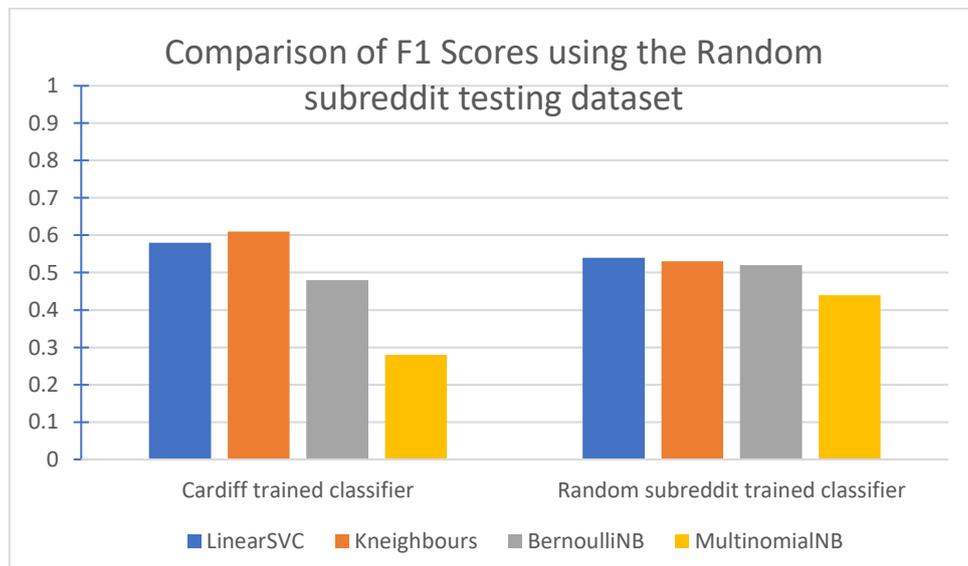


Figure 11: Graph of comparison of F1 scores from using the Random subreddit testing dataset with both types of classifiers

These outcomes could be for a number of reasons, one such possible reason is that there may have been a certain number of errors present in the data labelling process since I labelled the Cardiff dataset first and then the random subreddit data a number of weeks after that.

The time between labelling may have meant that I was much more confident in labelling something as narrative in the random subreddit dataset compared to the Cardiff one, which could explain why there is a larger number of narrative posts in the random subreddit dataset.

In order to determine if this were an outlier in my tests I would look into further attempts at re-labelling the datasets to see if there is a difference in the number of posts being labelled as narrative, and a further method to this would be to enlist the help of other testers to label the dataset to compare the results.

A further piece of analysis I have done in regard to this issue is to examine whether errors in my results were the results of the labelling strategy I chose; this has been explained further on in the evaluation where I examine a number of false positive and false negative results from the LinearSVC classifier.

Alternatively, another reason for this discrepancy could be that the types of narrative may differ between the two datasets, with many Cardiff posts being related to events that have been going on around the city during the past few days such as a post about nobody checking train tickets at a particular station, while the narratives in the random subreddit dataset are much more varied in nature from wider ranging topics such as discussions about video games and relationships.

Something of particular note about these datasets is that the Cardiff dataset possesses an average length of 529 characters for narrative posts while the random subreddit dataset has an average of 1211 characters for narrative posts.

Another goal that was outlined for this project was to determine what the best classifier was for this task. I found that the best results for the tests that I conducted were from using the RoBERTa model. Regardless of the test scenario RoBERTa performed the best.

The clear lead that the RoBERTa model has over the other classifiers was partially to be expected due to the strength of the training method, it being able to reach the top 15 for the GLUE benchmark leader board (Wang, et al., 2019) for NLP tasks. GLUE is a General Language Understanding Evaluation benchmark which contains a collection of datasets that are used to train, evaluate, and analyse language models. Being in the top 15 puts RoBERTa above many other models and means it will likely have a very strong performance compared to simpler models like the LinearSVC classifier.

Compared to the other non-RoBERTa classifiers the second best result that I observed from my results was using the KNeighbours classifier in the test where I trained the classifier on the Cardiff dataset and tested it on the random subreddit dataset.

There was a significant difference between the two best results, with RoBERTa having an F1 score of 0.76 compared to the KNeighbours classifier having a score of 0.61.

While the KNeighbours classifier was able to achieve the second highest score across all other classifiers, it was not consistently higher than the other classifiers. On average the LinearSVC classifier was able to achieve the highest score if we discount the scores for RoBERTa.

This could simply be a fluke on the part of the KNeighbours classifier, and this could be resolved by doing additional tests to determine if the F1 score of 0.61 was an outlier.

The overall worst performing classifier was the multinomial naïve Bayes classifier, with an average score of 0.38. This does make sense as the multinomial classifier tends to perform worse when the dataset consists of shorter texts.

We can observe this to be true in the tests conducted, as the Cardiff dataset has a lower average character count compared to the random subreddit dataset. Figures 3 and 5 show that the Multinomial Naïve Bayes classifier performs better with the random subreddit dataset compared to the Cardiff dataset.

A further point of analysis is to determine where each classifier seems to be making mistakes. To evaluate this, I have selected 5 false positives and 5 false negatives for each LinearSVC classifier test and have tried to find areas where the classifiers appear to be having issues.

The following tables illustrate examples from the LinearSVC tests which are false positive, meaning that the classifier has incorrectly labelled a non-narrative post as a narrative, and false negative, where the classifier has incorrectly labelled a narrative post as non-narrative.

The tables will also contain the confidence score for each post, which is the signed value that shows how confident the classifier is that a particular post is narrative or non-narrative, with a positive value meaning that the classifier would predict a post to be non-narrative and a negative value meaning that the classifier would predict the post to be narrative, with a greater number meaning that the classifier is more confident in its decision.

Table 1: Table of 5 False positive and 5 false negative results using the Cardiff LinearSVC classifier tested on Cardiff data

False Positives		False Negatives	
<i>Post</i>	<i>Confidence score</i>	<i>Post</i>	<i>Confidence score</i>
I'm alone in town until tomorrow, monday, who's up for a beer or some food or anything else? I'm a 25 year old guy Thanks!	-0.185	I'm turning 18 soon and was thinking about how I can finally order alcohol at a pub/bar legally . Problem is, a lot of the bars/pubs I've been to had an older crowd (no offense to the older generation :P just prefer a younger crowd) and wanted to know if there's any place that students and people aged 18-26 frequent a lot. Not sure about the Covid situation atm and regulations at the moment, so this is mostly just for future reference. Thanks :-)	0.022
I've heard that most clubs have scanners for IDs now. Does this scanner accept something like a driving licence or does it have to be a specific type of ID? Google doesn't seem to come up with any results besides a BBC	-0.984	I have curly / wavy hair and in desperate need for a haircut. I used to go to a place up north that would dry cut my hair and then wash after and it would come out perfect. They specialised in curly hair. I've tried and failed to find a place in Cardiff that offers a hair cut without washing my	0.376

article on the scanners in Northern Ireland so I thought I'd ask here.		hair before the cut. Because of the nature of my curly hair, when wet, its super straight, so when it is cut this way, it dries and the curls can look really odd and uneven. Can anyone provide a place in Cardiff that offers this ??	
Can you bike through them, or are they still completely flooded?	-2.399	There's a load of police blocking off an area of town. Anyone know what's going on there?	1.786
Anyone noticed there's a lot more coppers on the beat lately? Are they more obvious or are they checking on pubs etc being shut?	-1.018	Especially around Lloyd George Ave and town. Seems to be a load of coaches causing a rucus.	0.236
I've been invited to an 'Experience Day' but I was just wondering if anyone here had some first hand experience with the place	-2.535	Anyone else find it a bit concerning how much the group has grown since that tiny group protested kickdown initially? Makes me a bit nervous about how much of the anti intellectualism has leaked over from the states.	0.337

The highest confidence score where the classifier predicted a narrative correctly was -6.4 with an average score of -1.7, and the score for predicting a non-narrative correctly was 8.2 with an average score of 2.6, this provides us with a scale for how confident the classifier is for certain posts.

One aspect where I believe the classifier struggles is with the writing style of certain posts such as post 3 from the false positive section and post 4 from the false negative section. These posts can be understood to be non-narrative and narrative by a human due to the implied context of the post, which is what I have used to determine the label. Particularly post 4 because the person writing the post has missed out a lot of context of the story that they are telling involving an event of there being many coaches around Lloyd George avenue, the classifier is unable to determine that the post is a narrative because it does not understand the minimal amount of information given.

Another issue with labelling that is seen from table 1 is that there are cases, such as post 2 from the false negatives section, where I have labelled a post as either a narrative or non-narrative where it could be argued that the post is both a narrative and not a narrative in different parts and this confuses the classifier as it has a score of 0.3 meaning it is almost on the edge of considering the post either narrative or non-narrative.

A solution that I considered to this problem is to label the posts using a scale from narrative to non-narrative as mentioned previously. This is something that I considered doing but instead opted to label the posts as a binary yes or no as it would make the tests much simpler. Another solution that I considered was to label posts at the sentence level, this would avoid many of the issues that I had in the labelling process where I wasn't sure what to label an overall post.

An additional problem in table 1 is there are examples where I have labelled some posts incorrectly, for example post 1 and 5 in the false negative column, this is an expected outcome from labelling 1000s of posts in a few sittings and would be mitigated by having more people labelling posts and comparing results with each other.

A possibility for post 5 in the false positives section being labelled incorrectly is that the classifier is being confused by the typos that the person writing the post has put in. The post could be interpreted as a narrative by taking it at face value, however as a human I was able to understand the intention behind the post as being more non-narrative focused while the classifier has deemed the post to very likely be a narrative with a score of -2.5.

Table 2: Table of 5 false positive and 5 false negative results using the Cardiff LinearSVC classifier tested on random subreddit data

False Positives		False Negatives	
<i>Post</i>	<i>Confidence score</i>	<i>Post</i>	<i>Confidence score</i>
I know I am it's gonna be the first thing I unlock. I'm gonna laughing when I'm kicking those captains, bosses and mercenaries off ledges/hills and Just watching their health deplete a ton.	-3.151	Hi all. On previous note variants you could push the button on the S pen and highlight the text in any application. Much like a mouse on a computer. I tried to copy something from Facebook last night and i could not do it. It works on Google but not is apps. Any ideas on why?	1.022
I dont think it would get detected by YT under this name; also gotta put "NOT ASSOCIATED WITH HAMPTON BRANDON" in the channel description and video descriptions to save face I would do this myself but I dont have the time or resources to properly do this (i also dont know how to do it kek)	-1.186	Ive seen pictures and videos of people with these crops that take up a 3 x 3 space. How do you get them?	1.714
It's a show about how corporations leverage their power to force innocent people into compromising positions, and immediately discards them thereafter. There are subplots about the US	-1.883	If Im trying to check out a song, and I hear right of the bat its a piano intro, I turn it off and forget about it. People think they sound sweet and beautiful for some reason, but they are really easy to create	0.613

<p>meddling in the politics of SE Asia, fascist evangelicals, and corporations cynically co-opting the Me Too movement. It's an over budgeted superhero show that streams on Amazon Prime.</p>		<p>and often dont even sound good. It is very rare for me to hear a piano-led song that I like. I cant even think of one as I type this.</p>	
<p>I'm looking to start pressing my own shadows. I've been comparing different high-end brand ingredients and they all seem to be a bit overwhelming. Before I continue ordering all the ingredients I was wondering if anyone had a good recipe for good soft matte shadows with light fallout. Most of what I'm ordering is from TKB. So far I have TKB Matte Texture Base; Zinc Stearate; MyMix Clear Pressing Binder Medium; and the pigments. Help me out my fellow at home chemists!</p>	-0.143	<p>There was a post last night that was a screenshot from Netflix. It had the premiere date as November 9. I looked for it today to check for more updates and its no where to be found. Was it fake?</p>	0.706
<p>Any recommendations for computers that can handle pretty demanding games (ARMA 3, Insurgency: Sandstorm, etc...) for under \$1000?</p>	-2.042	<p>What would be the consequences of consuming around 40mg of Cyclobenaprine and a 1/5 of Vodka? After doing some research, I've read the a majority of people end up sleeping for a couple days but eventually wake up. Is this true? thanks.</p>	2.604

The highest confidence score where the classifier predicted a narrative correctly was -3.5 with an average score of -1.4, and the score for predicting a non-narrative correctly was 7.7 with an average score of 2.4.

Overall the scores for testing the Cardiff LinearSVC classifier on random subreddit data are similar to that of testing on Cardiff data, with minor differences in average score.

Something that was present in the previous table that is the same as in this one is that there are more cases where there are posts that could be labelled using a scale from narrative to non-narrative. This appears to be an area where the dataset falls short as many posts contain a mixture of contents that the classifier seems to struggle with.

Something that interested me about post 4 from the false negatives section was to consider how much weight the final sentence gives to the classifier for it to classify the post as non-narrative. I performed a test and found that simply removing the question mark resulted in a 0.3 score difference, which means that the classifier seems to think that non-narrative posts are more likely to contain question marks than narratives.

I investigated this further using post 5 from the false negatives section and found the same results, simply removing the question marks in the post resulted in an improvement in confidence score from 2.6 to 1.7.

This problem indicates that the classifiers are understanding aspects of the text that make up a narrative, however it could simply be that the posts that I have labelled as non-narrative appear to contain question marks more often than posts which are labelled as narrative.

One issue I foresaw with some of these posts is the lack of consistency with the writing styles and general language used. Particularly in post 3 of the false negatives section the person writing the post has missed an apostrophe when writing the word "I'm", when I tested this same post with the apostrophe included the classifier gave it a score of 0.3 instead of 0.6.

This shows that spelling and punctuation errors are playing a role in the decision making process of the classifiers. This issue could also be an outlier based on the dataset that I have gathered, since Reddit does not require users to post grammatically correct posts it may simply be that there were more spelling errors in the non-narrative subset of posts compared to the narratives.

This is a problem that could be solved by manually proofreading each post, however this would take a large amount of time and was not feasible for a project of this scale.

Table 3: Table of 5 false positive and 5 false negative results using the random subreddit LinearSVC classifier tested on random subreddit data

False Positives		False Negatives	
<i>Post</i>	<i>Confidence score</i>	<i>Post</i>	<i>Confidence score</i>
A boss killed me and i cant seem to find my blood echoes. Havent gone in again to check i know he will f**k me up. Also do bold	-0.28267	Hi guys, sorry if this question pops up a lot. I'm hoping you can help! Over the past few months my boxers and boxer briefs continue to ride	0.609518

<p>hunter marks work in boss battles?</p>		<p>up and get really uncomfortable. I'm considering switching to briefs to avoid it. Anyone else experience this? Is there a comfy brand of underwear (for briefs, boxer briefs, or boxers) that is worth switching to that also has a bit of space up front? Thanks for your help!</p>	
<p>I'm a direct descendant of Isaac Galland. He is known for selling land to Joseph Smith in what is known as Navuoo. There's some other stuff he did and my mother had his journals. She gave them over to BYU. Fun fact he was also Joesph Smith's secretary and in the end figured him to be a fraud.</p>	<p>-0.2774</p>	<p>Hi all. On previous note variants you could push the button on the S pen and highlight the text in any application. Much like a mouse on a computer. I tried to copy something from Facebook last night and i could not do it. It works on Google but not is apps. Any ideas on why?</p>	<p>0.78539</p>
<p>Have a whole bunch of these guys popping up from some mulch under my orange tree in Houston.</p>	<p>-0.24203</p>	<p>Really stressing out because I'm not sure if I got scammed or what. Basically sold something for 200 on ebay, got an offer for that much and printed off the label and sent it away. I sent that away on the 19th but nothing has come up in my paypal that I have the payment received but everything on ebay is saying that everything has went through alright.</p>	<p>0.10905</p>
<p>Havent played with in a while. Has she been fixed? I recall she had</p>	<p>-1.62611</p>	<p>I get a text message. Watch vibrates, phone does not (and is not</p>	<p>0.923576</p>

this weird thing were she wouldnt dash to an enemy		on silent mode). Didnt do this prior to getting iPhone 11. Thoughts?	
Some update your ai. I didn't write that. We lifted sanctions and they did nothing at all. Previously the US said do x and we do y. Trump said do something and we do y. They did nothing. An incredibly bad negotiation if there was one. /u/whatdc	-0.24142	I tried to claim penthouse for gta v online on twitch prime but it stuck with the loading thingy everytime i do that, instead of claimed!	0.939272

The highest confidence score where the classifier predicted a narrative correctly was -1.1 with an average score of -0.4, and the score for predicting a non-narrative correctly was 7.3 with an average score of 1.0.

The scores for the random subreddit classifier are much lower compared to those of the Cardiff classifier seen in tables 1 and 2, with an average narrative score of -0.4 compared to -1.4 and -1.7. This indicates that the random subreddit data may be less clear about what a narrative is compared to the Cardiff data.

A common theme among the labelling strategy that I have chosen to use is that there are many posts which contain both narrative and non-narrative elements, for example post 1 in the false negatives section.

To test this, I separated the post into the parts that I consider narrative and non-narrative. By simply separating out the individual sections of the post the classifier is able to identify the different sections correctly. I gave the following sections to the classifier:

```
"Hi guys, sorry if this question pops up a lot. I'm hoping you can help!",

"Over the past few months my boxers and boxer briefs continue to ride up and get really uncomfortable.",

"I'm considering switching to briefs to avoid it. Anyone else experience this? Is there a comfy brand of underwear (for briefs, boxer briefs, or boxers) that is worth switching to that also has a bit of space up front? Thanks for your help!"]
```

Figure 12: Image of text separation from post 1 in the false negatives section of table 3

After isolating the parts of the post that I considered to be narrative and non-narrative the results were surprisingly good. With a score of 1.2 for the first section, -0.7 for the second, and 0.9 for the third, the classifier correctly labels the individual sections correctly but struggles when they are combined into one post.

This information suggests that opting to label posts using a scale from narrative to non-narrative would likely have been the better choice, as it is rare that a post ends up being entirely narrative or entirely non-narrative.

One post that I particularly found to be interesting was post 3 from the false negatives section. The classifier predicts this post to be non-narrative by a score of only 0.1, meaning that it is right on the edge of being correctly labelled.

I performed a test by editing the text to see what would get the classifier to label this post as a narrative. My first idea was to remove the initial sentence, as I considered this part to be non-narrative, which resulted in a score of 0.04.

"Basically sold something for 200 on ebay, got an offer for that much and printed off the label and sent it away. I sent that away on the 19th but nothing has come up in my paypal that I have the payment received but everything on ebay is saying that everything has went through alright.",

Figure 13: Image of test text used to compute decision scores from post 3 in the false negatives section of table 3

While this was an improvement in score it still labels the post as non-narrative by a very small margin, my next idea was to make the post grammatically correct by adding "I" to places where it is needed in the post.

"I sold something for 200 on ebay, I got an offer for that much and printed off the label and sent it away. I sent that away on the 19th but nothing has come up in my paypal that I have the payment received but everything on ebay is saying that everything has went through alright."

Figure 14: Image of test text used to compute decision scores from post 3 in the false negatives section of table 3

This resulted in a successful labelling of the post as a narrative with a score of -0.02, this indicates that the classifier places some degree of weight on the use of "I" in narrative posts compared to non-narratives. If this post were grammatically correct it would be more likely to be labelled as a narrative due to the increase in the amount of identifiable narrative aspects which were otherwise missed due to the spelling mistakes.

Table 4: Table of 5 false positive and 5 false negative results using the random subreddit LinearSVC classifier tested on Cardiff data

False Positives		False Negatives	
Post	Confidence score	Post	Confidence score
Been in London for a few months now, and honestly so sick of this city just to sum up, the crowds, daily depressing commute, ratrace, smell, and mainly the lack of escape to anywhere I can feel all alone. I	-0.05629	So posted a few months ago about coming out as transgender etc and yeah kinda a mess atm like minimal friends feeling like I'm stuck tbh and also does anyone know if there's anywhere I can	0.265473

<p>love having a bit of a beach/forest/countrys ide where you can escape and just feel at peace when you need to nearby. Considering Cardiff as i've been down there a few times and really love how close it is to nature, surfing spots, beautiful south west etc. But I need some advice on the city and pros and cons of living there?</p>		<p>request a clothing donation new wardrobe is needed I hate the masculine clothes I have me and shops don't go atm out of comfort zone and shops that are open rarely do my size</p>	
<p>I was just wondering tbh.</p>	<p>-0.55276</p>	<p>On Friday evening (30/08) I lost my brown key pouch with a car key (Vauxhall), key fob from my apartment, and few other keys. I lost it in the area of the Dock pub in Cardiff Bay. If anyone found the pouch or a car key with Vauxhall logo on it, or any other keys nearby PLEASE get in touch.</p>	<p>0.241612</p>
<p>Tried ringing three Halfords but no answer.</p>	<p>-0.56726</p>	<p>Hi guys, At Car Free Day in town yesterday, I found a phone (an older iPhone) with a distinctive wood effect cover and student card for a Louis Reynolds. Phone is dead so can't try switching on and the student card says 'SCG' on it but I've no idea what that means. Long shot but does anyone know either this person or what SCG means? Thanks!</p>	<p>1.51</p>

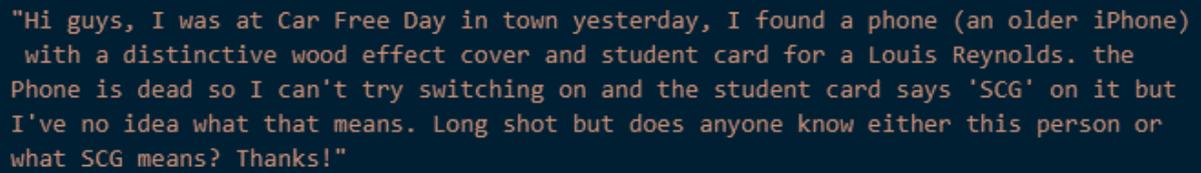
Looking for hall, bedroom, lounge, and been given a bit of a run around.	-0.48263	Today I was hoping to get to Barry island asap. They have built a platform specifacly for Barry. Turns out they only have 2 coaches on a day that's meant to get to 30C and in the summer holidays. The council really needs to take notes of these things and do something about. When the train arrived it was already full and no space to stand up. This needs to change	0.379417
Haven't been able to attend the other ones but was curious about them	-0.42534	I've developed an awful phobia, thanataphobia. Fear of death. It's consuming tbh. I'm looking for a counsellor or therapist or some sort of appropriate health care professional that can help with this sorta thing. Thanks.	0.303607

The highest confidence score where the classifier predicted a narrative correctly was -1.2 with an average score of -0.3, and the score for predicting a non-narrative correctly was 3.6 with an average score of 1.0.

The scores for testing the random subreddit LinearSVC classifier on Cardiff data are very similar to those when testing on the random subreddit data, with only a minor 0.1 difference in average score for classifying narrative posts correctly.

From table 4 one of the posts which interests me the most is post 3 from the false negatives section, with a score of 1.51 this was the worst labelled narrative post from the entire test. I found this to be confusing as I feel that the post is generally much more narrative than it is non-narrative. I suspected that this was possibly to do with the shorthand nature of how the writer has structured the post, with many words being missed out and implied which would confuse the classifier but be understandable to a human.

To test this I rewrote parts of the post to see if it would affect the score being given by the classifier.



"Hi guys, I was at Car Free Day in town yesterday, I found a phone (an older iPhone) with a distinctive wood effect cover and student card for a Louis Reynolds. the Phone is dead so I can't try switching on and the student card says 'SCG' on it but I've no idea what that means. Long shot but does anyone know either this person or what SCG means? Thanks!"

Figure 15: Image of test text from post 3 in table 4 showing the addition of a few words

The above figure shows the new text, which when classified by the classifier provides an improvement in score of 0.3 points, showing that while the improvement was small it certainly was affecting how the classifier was interpreting the data.

This is mostly due to the addition of the pronoun "I" which the classifier considers is making the post more narrative, however the score is still very high, and I believe this is because the classifier interprets more sections of the text as non-narrative than it does narrative. Again this is a case where a scaled labelling method would be beneficial to see more in-depth results.

This error analysis has provided me with a number of interesting conclusions about the data that I have gathered and labelled. One of the primary outcomes from this analysis is the understanding that the classifiers appear to be assigning a very large degree of importance in the use of question marks in non-narrative posts, indicating that the use of question marks may be incorrectly influencing the classifiers to skew a post towards a non-narrative label.

Another outcome of this analysis is the understanding that personal pronouns such as "I" are also influencing the classifiers to label posts as narrative.

An overall outcome that I observed from this error analysis is that I as the labeller of the datasets have made a number of mistakes in my labelling process by incorrectly labelling posts. An outcome that is linked to this is the understanding that the labelling strategy I have used of giving posts an overall label has been detrimental in a number of posts in my datasets, and this has led me to propose alternate strategies that would provide clearer results such as using a scaled labelling system or sentence level labelling.

Chapter 5: Future work

If the scope of this project were larger I would definitely consider a number of ideas to expand on the evaluations I have already made.

There are a number of improvements to be made on the datasets that I have already gathered. One point of contention is the size of the datasets, while 1000 posts per category was a reasonable goal to aim for given the size of this project, the conclusions that I could make on this data would be much better if I could expand the sizes of the datasets to perhaps 10000 posts. Since having more training data would allow for more types of posts to be given to the classifiers to better build an understanding of the types of narratives, it would also provide a clearer image of what the classifiers are understanding from the data as more data means that there are more possibilities for the classifiers to understand the contents of posts.

Another point to improve on with the dataset is to expand the categories to other topics, for example instead of selecting from random subreddits I would instead choose to select specific topics such as sports, games, literature, or science. By doing this I would be able to identify if certain topics provided better sources of narrative posts compared to others.

While evaluating my data I found that there were areas where the classifiers would struggle, particularly when posts had spelling and grammar errors. A possible method of solving this issue would be to proofread the datasets that have been gathered, however this would overall be very difficult to do if the number of posts is scaled up. As such the issue of spelling errors is likely unsolvable for a dataset of sufficient size.

Further to gathering additional data, I would also change the labelling strategy that I used. While labelling the data by myself was a reasonable approach for a one-person project it would be better for a team of labellers to corroborate their labelling to determine as a team which posts should be labelled as a narrative.

Another way of improving the labelling process that I would do if I had more time is to use a scale for labelling posts as narrative vs non-narrative, as this could improve the classification of posts by allowing classifiers to understand which posts are more narrative than others.

Alternatively, separating each post down to the sentence level and labelling them instead of the overall post may prove to be a good method as I have found in my error analysis that posts may be labelled incorrectly overall, but are labelled correctly when separated into each sentence.

To further test the data that I had, I would also suggest using the k-fold cross validation technique for the tests. As this method would have taken a great deal of extra time to implement and run I did not use it in this project, however if this project were to be expanded it may result in a much more accurate set of results for each test.

Chapter 6: Conclusions

In conclusion, this project began with a number of aims of comparing the effectiveness of using different sets of Reddit data to train language models to determine if a specific Reddit post was a narrative or not. With further goals of observing if these language models could understand what aspects of a Reddit post make it a narrative.

Overall, all of the goals set out for this project were completed on time.

Initially, this project began with the goal of gathering and labelling a Reddit dataset, which I completed by gathering over 1000 posts from the Cardiff subreddit and 1000 posts from a random selection of subreddits. I labelled these posts with binary values of narrative or non-narrative.

One of the downfalls of this project was my choice of labelling the datasets using a binary system. This was an issue because some posts could include elements of both narrative and non-narrative text. Given this, I have proposed alternative labelling methods that may improve on the results that I have gathered. These alternative strategies include, using a scaled labelling system between narrative and non-narrative, and a method of separating posts down to the sentence level and labelling the sentences individually. I found that these alternative methods would likely result in an increase in performance for the language models that I used.

One of the aims of this project was to determine which of the tested classifiers performed the best, and it was found that the RoBERTa language model certainly provided the best results across all the tested models, with the LinearSVC classifier coming in at a close second.

This project was undertaken to determine if Reddit posts could make a good source of narrative text, I believe that this project has determined that with a good enough filtering system, such as a narrative trained RoBERTa language model, Reddit can absolutely be a reliable source of narrative posts.

One of the overarching aims of this project was to determine if using a narrative classifier was a feasible method of filtering a Reddit corpus. After completing this project I believe that the use of the RoBERTa model would provide reasonable performance in filtering out non-narrative posts. However further improvements could be made in improving performance, such as the alternative labelling methods that I have mentioned above.

There are certainly issues involved with using language models other than RoBERTa, such as the LinearSVC classifier. These classifiers have drawn incorrect conclusions about what makes up a narrative, such as the use of question marks being much more likely to skew the classifier to label a post as a non-narrative.

The results of this project found interesting comparisons in the differences between using a specialized dataset compared to a more generalized one, in the test cases where I have tested

While this project has certainly provided interesting results, I believe it has more so accomplished setting a baseline for the further research into the topic of using Reddit as a dataset for natural language processing tasks.

Chapter 7: Reflection on learning

I believe my approach to this project had been partly marred by the ongoing coronavirus pandemic, while I was able to progress relatively within the planned time constraints that I set myself, I still felt that I could not spend as much time on this project as I would have liked to due to stress.

My initial plan timeline mostly matched how the project developed, with the only missed milestone being finishing the implementation by week 8. The easter break was used as extra development time in order to be able to start writing the report by week 9.

While implementing the classifiers using Scikit-learn was a relatively painless experience, the implementation of RoBERTa was much harder. While I am happy that I eventually managed to get the algorithm working it was certainly one of the hardest challenges that I faced during this project, mostly due to the hardware and software issues as well as a lack of official documentation for the Simple Transformers library.

I can now say that this project has allowed me to become much more confident in my ability to implement any of the Scikit-learn classifiers in the future. If I am required to do any natural language processing tasks or research in the future that uses Scikit-learn or the Simple Transformers library, I will certainly be much more comfortable implementing them thanks to the experience I have gained in this project.

The section of the project that I most enjoyed was the data gathering and labelling aspects. As someone who uses Reddit on a daily basis, it was very interesting to go through individual posts and read the stories within them.

Overall this project has challenged me in a number of ways that I have not been challenged before and has forced me to develop new skills in areas where I previously had none. It has been an exciting opportunity to delve into the topic of natural language processing and has provided me with an interest in the topic for the future.

References

- Crews, F. C., 1977. The Random House handbook. In: *The Random House handbook*. New York: Random House, p. 14.
- Facebook AI, 2019. *RoBERTa: An optimized method for pretraining self-supervised NLP systems*. [Online]
Available at: <https://ai.facebook.com/blog/roberta-an-optimized-method-for-pretraining-self-supervised-nlp-systems/>
[Accessed 24 April 2021].
- Goldberger, J., Roweis, S., Hinton, G. & Salakhutdinov, R., 2005. Neighbourhood Components Analysis. In: *Advances in Neural Information Processing Systems*. s.l.:s.n., p. 513.
- Gordon, A. S. & Swanson, R., 2009. Identifying Personal Stories in Millions of Weblog Entries. In: *Identifying Personal Stories in Millions of Weblog Entries*. Los Angeles, California: s.n., pp. 17-18.
- Jurafsky, D., Martin H., J. & Kehler, A., 2020. *Speech and language processing : an introduction to natural language processing, computational linguistics, and speech recognition*. 3rd ed. Upper Saddle River, NJ: Prentice Hall.
- Levengood, C., 2020. *Building a custom Scikit-learn Transformer using GloVe vectors from Spacy as features*. [Online]
Available at: <https://lvngd.com/blog/spacy-word-vectors-as-features-in-scikit-learn/>
[Accessed 10 March 2021].
- McCallum, A. & Nigam, K., 1998. A comparison of event models for Naive Bayes text classification. *AAAI-98 Workshop on learning for text categorization*, p. 42.
- Pedregosa, F. et al., 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, Volume 12, pp. 2825-2830.
- Rajapakse, T., 2020. *About*. [Online]
Available at: <https://simpletransformers.ai/about/#credits>
[Accessed 26 April 2021].
- Scikit-learn, n.d. *Classification of text documents using sparse features*. [Online]
Available at: https://scikit-learn.org/stable/auto_examples/text/plot_document_classification_20newsgroups.html
[Accessed 11 March 2021].
- Spacy, 2020. *Spacy*. [Online]
Available at: <https://spacy.io/usage/facts-figures>
[Accessed 26 April 2021].
- Wang, A. et al., 2019. *Glue Benchmark Leaderboard*. [Online]
Available at: <https://gluebenchmark.com/leaderboard>
[Accessed 02 May 2021].