Dissertation Type: Applied

DEPARTMENT OF ENGINEERING MATHEMATICS

An analysis of neologistic derogatory terms

and their uses within Common Crawl

Harrison Bennion

A dissertation submitted to the University of Bristol in accordance with the requirements of the degree of Master of Science in the Faculty of Engineering.

Monday 12th September, 2022

Supervisor: Dr. Matthew Edwards

Declaration

This dissertation is submitted to the University of Bristol in accordance with the requirements of the degree of MSc in the Faculty of Engineering. It has not been submitted for any other degree or diploma of any examining body. Except where specifically acknowledged, it is all the work of the Author.

Harrison Bennion, Monday 12th September, 2022

Contents

List of Figures

List of Tables

Ethics Statement

An ethics application for this project was reviewed and approved by the faculty research ethics committee as application 10820.

Abstract

This dissertation explores the creation and use of neoligistic slurs throughout the internet through the use of Urban Dictionary and Common Crawl. Work by Luccioni and Viviano, 2021 [18]) investigated models making use of Common Crawl as a corpus for training models, discovering that many of the models contain a large amount of undesirable content, which has an impact on language models and can lead to biases and hateful pejoratives being present within their embeddings. By analysing the unusual uses of slurs within Common Crawl, it's possible to gain a much better understanding of how these can be monitored and accounted for when training models.

Wilson *et al.*, 2020, [26] and Nguyen *et al.*, 2018, [21] have previously investigated the use of Urban Dictionary as an informal corpus for language modeling. They found the website contained a high amount of crude and explicit content, which is difficult to filter out. However, this paper makes use of these definitions by collecting a list of slurs from Urban Dictionary and utilising a participant survey, filtering these down to the unusual terms before exploring their usage within Common Crawl.

This method found a number of websites and forums where hateful language was being used through KMeans clustering and VADER sentiment analysis models. Based on additional research, it was possible to detect a few instances of reclamatory insults being used too.

A summary of achievements:

- 100 hours spent collecting and researching material relating to the current state of social media abusive language detection, undesirable content within natural language processing models, and the use of neologisms within communities online
- 536 lines of *Python* code was written using Jupyter notebook and Pycharm which are included in the repository
- Extracted derogatory and insulting terms from Urban Dictionary dataset
- Classified unusual terms into two ranges using a participant survey
- Identified usages of neologistic insults being used in offensive manner within the Common Crawl dataset
- Discovered certain terms being used in a reclamatory sense

Supporting Technologies

Below is a summary of the technologies used within this project

- I used *pandas*, *numpy*, *time*, and *random* public-domian Python Libraries for data-handling and basic processes.
- I used *jupyter notebook* public-domain Python Library and *PyCharm* Community Edition as an editing tool to develop the codebase.
- *Conda* was used as an open-source environment management tool. An environment file will be included within the project's files to allow others to emulate similar conditions.
- *pandas* and *matplotlib* were used to visualise outputs from the various programs.
- *enchant*, *re*, *warcio*, *multiprocessing* and *requests* are public-domain Python libraries which were used for processing the data and building the models.
- Parts of *gensim*, *sklearn*, and *vaderSentiment* were used for advanced modelling and processing, for example KMeans and Word2Vec models.
- Amazon Web Services Athena and S3 were used to store and trawl the Common Crawl Index files
- I used LaTeX to format my dissertation, via the online service *Overleaf*.

Notation and Acronyms

UD-20	:	Urban Dictionary crawl from July 2020
CC	:	Common Crawl
WARC	:	Web ARChive
WAT	:	WARC metadata
WET	:	WARC text data extract
BERT	:	Bidirectional Encoder Representations from Transformers
TF-IDF	:	Term Frequency-Inverse Document Frequency
LASER	:	Language-Agnostic SEntence Representations
LR	:	Logistic Regression
VADER	:	Valence Aware Dictionary and sEntiment Reasoner
v_n	:	Votes for a given word, n.
i_n	:	Interaction for a given word, n.

Acknowledgements

I would first like to thank my supervisor, Dr Matthew Edwards. His continued assistance has been a great help in completing this piece of work.

I would also like to personally thank the many participants who were a part of my survey.

As a final note of acknowledgement, I would like to thank my family for their continuous support during my studies over the last few years.

Chapter 1

Introduction

1.1 Project overview

The world contains a huge number of people who wish to degrade and insult each other due to their race, condition, sexuality, gender or other characteristics. With the dissemination of the internet, wretched hives of scum and villainy have cropped up where these like-minded individuals can collate together in echo-chambers of their own mantra. Within these forums and communities online, it's common to find niche terms and neologisms being used as people continue to create new and inventive terms to convey ideas or themes that may have never been uttered before. But not every evolution of our lexicon is to convey a new subject or idea; sometimes it's just to insult someone or something. These new insults and slurs are constantly being coined as individuals require new phrases and terms to hit just a little harder, whilst also allowing them to communicate with other forum members discreetly.

The work completed here aims to assist in detecting shifts in behaviour and mood through analysis of hateful terms, including their creation and dissemination through various networks. In addition to this, being able to increase detection of abusive communities or individuals could assist website owners if they are trying to cease certain behaviour or prevent forums from becoming toxic.

1.2 Urban Dictionary dataset

As described by Wilson *et al.*, 2020 [26], "Urban Dictionary is a crowd-sourced dictionary for (mostly) English-language terms or definitions that are not typically captured by traditional dictionaries. In the best cases, users provide definitions for new and emerging language, while in reality, many entries are a mix of honest definitions, jokes, personal messages, and inappropriate or offensive language (Nguyen *et al.*, 2018) [21]."

The Urban Dictionary dataset being used within this project has no published source and was instead provided to me by Dr Matthew Edwards[1] at the onset of this work. The data was scraped directly from the Urban Dictionary website on 13/07/2020 with definitions ranging from this date back to the site's origins in 1999. This dataset will be referred to as UD-20 for the remainder of this paper.

The dataset consists of one jsonl file which contains 3,857,168 definitions of various terms. The data strucutre is defined below in table 1.1.

There are other data sets containing a list of derogaotry terms like the English Derogatory Terms list on Wiktionary but these terms don't allow for user rating, which will be used to determine whether a term is defined as unusual or not. In addition to this, due to their heavier moderation, neologies, or terms used within smaller circles, may be left off from their website, whereas Urban Dictionary allows anyone to add to their site.

1.3 Common Crawl dataset

The common crawl[2] dataset is an open-source collection of archived web pages which are stored within AWS S3. These files are trawled automatically by bots every few months, creating terabytes of data which anyone can work with. For this project, the July 2020 crawl was selected as this matches up with

[1] OneDrive Link
[2] `https://commoncrawl.org`

Column name	Description
defid	Definition ID from Urban dictionary
row	Arbitrary index column
word	Definition word or phrase
meaning	Meaning of definition
example	Example of definition in use
contributor	Username of contributor
date	Datetime of definition creation
upvotes	Number of upvotes
downvotes	Number of downvotes
crawled_time	Datetime taken from Urban Dictionary
gif	Link to included gif file, or NaN

Table 1.1: UD-20 Columns

UD-20's time frame. The data provided by Common Crawl is vast, and so particular processing tools and algorithms will be required to process this problem, which will be discussed in greater detail later in this dissertation. For the remainder of this paper, the Common Crawl dataset will be referred to by the acronym CC.

1.4 Challenges

The effective use of the Common Crawl dataset will be the main challenge this project faces. Processing problems will result from both the data volume and the semi-structured nature of web pages. However, the Common Crawl website provides a great selection of tools and resources to help users access the data quickly, and many people have come up with innovative solutions to the issue. As this project concentrates on what is being discussed within websites, we will concentrate on the raw text data stored within the WET archive files, whereas many alternative projects primarily focus on WARC files and the URLs contained within Common Crawl.

Correctly identifying proper uses of the terms and phrases will be another challenge for this project, as many forums and websites will be studied, many of which will use these terms in differing manners. Additionally, because of the term's varied usage and definition within Common Crawl, it will be challenging to manage false positives.

1.5 Project Aims and Contributions

This project first aims to define what can be considered an unusual slur from the UD-20 dataset by using responses gathered by participant engagement. This classification aims to exclude well-known terms that most people know and use regularly, whilst also removing completely unknown terms.

After creating this definition, the next aim is to locate particular sites or forums where these terms are being used, with the follow-up aim of completing network analysis of these sites in order to locate the creations and uses of the neologistic terms.

As an extension of this idea, there is another aim to detect terms which have been reclaimed by their respective communities and are now being used in an ironic sense, as discussed by Cepollaro and López, 2020 [8]. In order to effectively detect whether terms are reclaimed or not, this project aims to use sentiment analysis to detect the meaning behind the terms being used.

1.6 Key achievements

A definition for textitunusual slurs was created by extracting a list of insults from the UD-20 dataset and having participants select known terms from a proportionate stratified sample, which, when combined with the voting system within UD-20, two ranges were created. Following on from this, the unusual terms were used in a search of the Common Crawl dataset, with occurances being recorded for processing.

By following up on the work by Luccioni and Viviano, 2021 [18], two sentiment analysis models were run on the CC dataset to identify hateful content within Common Crawl. From this, multiple

pages containing hateful content were found, whilst others were found to contain insults being used in a reclamatory manner.

Chapter 2

Context

Due to the range of topics covered within this project, this chapter aims to present context regarding the problems being faced and to discuss gaps in current research which this project seeks to assist with.

2.1 Derogatory and insulting language

In their simplest forms, slurs are typically defined as insulting or derogatory terms that can be applied to a particular group of people with the intention of shaming or degrading individuals. These terms have three differing forms of use as defined by Croom, 2013 [11], the paradigmatic derogatory use is the central use case in which they are used to insult a group of people. The second use is where slurs are targeted at those who are not part of their original intended target. For example, the use of *faggot* to refer to non-homosexual individuals, which is considered the non-paradigmatic use. Finally, there is the use of slurs in a non-derogatorily scene, where the term is used between "in-group" speakers as part of reclamation or in an ironic manner.

Reclaimed slurs are an interesting sub-group within the super-group of derogatory language. Terms are reclaimed for a number of different reasons, with a shift in social paradigms to allow in-groups and out-groups to form around them (Cepollaro and López, 2022 [8]).

In conjunction with the work by Croom, Martínez and Yus, 2013 [19], created a twenty-four case taxonomy to classify insults across any given cultural context. This work exemplifies the sheer range of contexts that insults may be used in, from conventional offensive tones to innovative terms used for social bonding. Although this taxonomy may be too meticulous, this approach could prove useful for the classification of terms within a larger model.

2.2 Neologisms

With the dissemination of the internet and the rapid development of technology, people have begun using the creative and flexible nature of language to coin terms to assist in conveying ideas, features, functions, and expressions with ease. An analysis of Netspeak Neologisms by Liu and Liu, 2014 [17], identified how groups used a number of methods like blending or giving old words new meanings to create neologisms as part of their dialect. In addition to this, it was found that most created terms tended to be noun compounds, with 93.5% of all compounds falling into this category, which further allows analysis to be fine tuned later in this paper. In addition to this, it was also noted that with the prevalence of the internet, linguistic diversity can be recorded at a much greater breadth and depth, which can then allow for much greater analysis.

Many communities have collated across the internet to discuss and celebrate various interests, hobbies, and ideas, lending further aid to the development of these communities and their mannerisms. Communities like the gaming community are continuously coining new terms as they are constantly adapting to new situations and ideas, as found by Belkova, 2018 [3]. Johnston, 2021 [16], found that a large portion of younger individuals understood these terms, with him speculating that media interaction types were a major factor.

However, not all of these communities that gather online are discussing hobbies or interests. There are a great number of forums and websites that host communities of foul-mouthed individuals. An article by Bogetić, 2022 [4], delves into one such outpost titled *incels*. Like many alt-right groups, the range

and speed of jargon produced is astounding, with Bogetić coining their language in a cryptolect of their own that almost escapes proper description. However, with their hateful attitude and *creative* use of language, a great many neologisms are slurs or insulting terms.

2.3 Undesirable Content within Datasets

In order to improve many natural language processing tasks, pre-training a language model has been shown to be highly effective (Dai and Le, 2015 [12]; Peters *et al.*, 2018 [24]; Radford *et al.*, 2018 [25]; Howard and Ruder, 2018 [15]). One such model that demonstrates this is BERT (Devlin *et al.*, 2019 [13]), which utilises a vast amount of training data. However, as studied by Luccioni and Viviano, 2021 [18], the data typically contains biases and undesirable content from the crevices of the internet, which can have an impact on embeddings and thus an impact on downstream decision-making. This study showed that there is a lot more work to do to uncover and filter out this content as many corpora, like Common Crawl, contain a great deal of it.

Chapter 3

Related Literature

3.1 Abusive language detection

When it comes to detecting abusive language, there are a great number of ways to do just this. One such method investigated by Pamungkas, Basile, and Patti, 2022 [22] analysed the role of swear words in abusive language detection. This paper found that modelling the task in a similar way to sentiment analysis led to promising results, with BERT-based models obtaining the best results. In addition to this, it was found that additional features like an 'abusiveness of the swear words' feature can improve detection, and so it may prove fruitful to create a form of this within this project.

There have been a few papers investigating general-purpose hate speech detection models, and one such interesting paper was written by Alaru *et al.*, 2020 [1]. This paper studied a number of models in a multilingual setting to study the generalisation potential of the models. It was found that BERT models tended to perform better at detecting abusive language when in a rich computing resource availability environment, whilst LASER [14] when used in conjunction with Logistic Regression (LR), was much more efficient in low-resource environments. BERT was found to search for context of the hate keywords, which was useful in context-heavy situations like on Twitter, whereas LASER + LR tended to just focus on the more hateful keywords as a much faster solution.

The previously mentioned work by Luccioni and Viviano, 2021 [18], also found that a variety of different approaches could be used to varying effects, with certain models favouring different aspects of derogatory language use. From this, it's clear that these models suffer from specialisation, as they may struggle to generalise to all problem spaces. Some of this issue may be present in the data space, as the internet sources favour those that use it, which is typically younger, English-speaking individuals from developed countries (The World Bank, 2020) [2].

3.1.1 Offensive language within Reddit

In addition to hate speech detection, there are a number of papers which study social media platforms and how it's possible to find nuanced offensive statements within them, which may not be immediately obvious at first. Breitfeller *et al.*, 2019 [6], investigated using sentiment analysis to detect microaggressions within Reddit. Their analysis found that many microaggressions are quite subjective, as there was a clear discrepancy between self-reported and observed microaggressions. Thus, detecting subtle uses of terms poses a distinct challenge. In addition to this, work by Morris, 2022 [20], found that whilst analysing swear word usage within Reddit, there were a great number of false-positive returns (swear words found in the wrong context), specifically within names and common terms. Alongside this, the use of extrapolated results could prove a highly useful method due to its reduction in computational power requirements in larger data sets.

3.1.2 Urban Dictionary

Wilson *et al.*, 2020a [27], studied the relationship between Twitter activity and Urban Dictionary, with events, popular figures, and memes trending on Twitter all having a direct impact on activity taking place on Urban Dictionary. Although this correlation is very general, it's a promising sign to see new terms being coined within Twitter quickly taking ground with Urban Dictionary. A paper by Nguyen *et al.*, 2018 [21], analysed the data within Urban Dictionary, finding many infrequent and informal words were

found within, but these quite often didn't correlate with known terms externally, as their crowd-sourced workers tended to only know terms with many definitions. However, because the purpose of this paper is to delve into the strange terms contained within, the data source should still be useful, even if the results will require some form of restriction before they can be used to find communities within Common Crawl.

Wilson *et al.*, 2020b, [26], attempted to study slang word embeddings using data from Urban Dictionary in another paper. By correcting the formatting of the terms and giving special treatment to multi-word embeddings, Wilson *et al.* found that the embeddings performed well in informal or non-literal language. Using a similar method to this, it could be possible to train embeddings using derogatory terms alongside a state-of-the-art pre-trained model to great efficiency.

Chapter 4

Design

This chapter of the report discusses the design choices behind the implementation of this work. The various sections follow the logical process of development from extracting the neologistic derogatory terms to analysing their use on the internet. As an overall principal, this work made use of GitHub and Gantt charts within Excel to maintain the project and control the versions of software during development. In terms of programming languages, Python was selected as the tool of choice due to its wide range of artificial intelligence and data handling libraries. Alongside, the ability to create virtual environments which can be easily re-created through their library capture process and the Jupyter Notebook functionality were key factors in selecting Python as the tool of use.

4.1 Overview of the approach

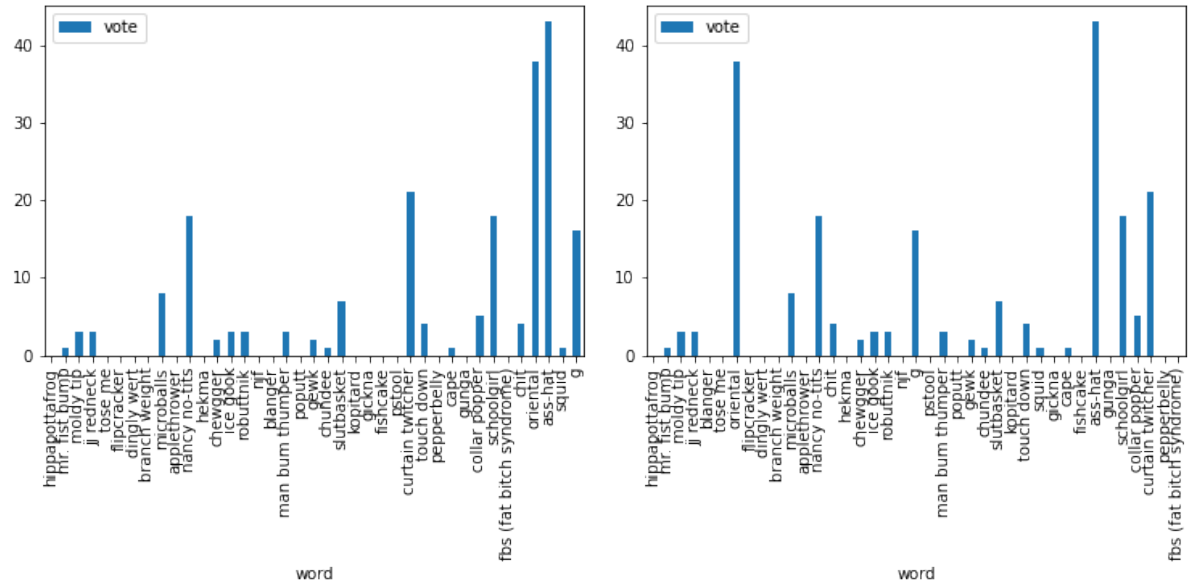
For this project, there were three primary sections of development. The codebase was produced iteratively, with a number of different approaches tested throughout. This chapter covers the final decisions in regards to the initial slur extraction from UD-20, the slur subset creation using the survey results, followed by the search of Common Crawl for uses of said terms, and finally the sentiment analysis on the found web page's text.

4.2 Neologisitic derogatory term extraction

4.2.1 Initial extraction and survey creation

The first challenge within this project was to extract a list of derogatory terms from the UD-20 dataset. In order to determine whether a term was unknown, unusual or well-known, it made sense to make use of the in-built metrics provided within Urban Dictionary. The voting system within the site allows anyone to vote on user-submitted definitions using up-votes and down-votes. Using these values, this paper derived two metrics; "interaction" is the total count of up-votes and down-votes, whereas the "score" is the up-votes minus the down-votes. By utilising these two values, we were able to rank and subsequently filter the list of terms.

A simple approach of regular expressions was used in order to refine the full UD-20 dataset down to just derogatory and insulting terms. Due to the lack of prior definitions or examples, a pre-trained model or neural network may have had issues in detecting fringe definitions that may have unusual language or targets. Instead, regular expressions can quickly search the various definitions for key phrases. A list of insults was compiled in order to determine whether a term is well-known or not. A participant survey was then used to gather human input to determine a cut-off point for the list. Prior to creating the survey, in order to determine whether a term should be included, the interaction score was focused on as this represented the amount of contact the definition had come into during its life on the website. From this, a stratified sample was used to select terms at regular intervals for the survey, allowing for a static dataset that all users would complete, whereas a random sample would cause additional issues with participant bias.



(a) Votes by word ordered by interaction

(b) Votes by word ordered by score

Figure 4.1: Comparison of distribution of votes

4.2.2 Defining unusual terms

Following on from the survey, it's possible to make use of the participants' votes to determine the bounds of an unusual term. To do this, it made sense to make use of the mean and fractions of the standard deviation of the votes to create subsets of slurs. From this, it was possible to define an upper and lower bound for the interaction value, which could then be used on the larger set of slurs. The votes were also compared to the score, but the distribution of votes showed a much weaker correlation, as shown in figure 4.1.

For this study, two subsets of slurs were created to compare the definition of unusual and to study how different levels of interaction can have on the Internet's usage.

4.2.3 Refining the list of slurs

From anecdotal evidence collated from the survey, it was clear that many of the terms people identified as knowing were in fact only known from their standard dictionary definition rather than their use as an insult, and so it made sense to only use the definition with the highest score from UD-20 moving forward. In addition to this, after the creation of the slur subsets and the beginning of analysis within Common Crawl, it quickly became apparent that the subsets still contained many false positive matches due to the community-driven nature of the definitions. To address this, it was clear that the poorly received definitions with a poor score to interaction ratio should be removed. Alongside this, terms made up of either single or pairs of characters, like *G*, tended to produce a high number of false matches, and so these were removed as well, with terms consisting of more than three words producing no-matches due to their extremely niche nature. The final refinement focused on the subset with a higher interaction band, as many of the matches from the subset were common English terms like *Cant*, and so terms that appeared in the dictionary were removed too in order to prevent more false matches.

4.3 Common Crawl subset and extraction

4.3.1 Understanding Common Crawl data structure

Within the Common Crawl corpus, there are a number of different crawls which are completed every few months by their automated bots. For this particular paper, it makes sense to work with the CC-MAIN-2020-29 crawl as this corresponds to July 2020, the same as the UD-20 dataset creation month. Within the crawl, there are 60,000 segments and files, which hold multiple websites' information. The crawl information is stored in three different file formats; WARC, which is the raw crawl data; WAT,

which holds the computed metadata of the WARC; and finally the WET files, which contain the crawled plaintext. For this task, it makes sense to use the WET files as the data source due to the minimal additional data processing required to make use of their information.

4.3.2 Extracting text samples

In order for a WET file to be considered for analysis, this project made sure to only include pages that had been identified as English by the crawling bots. Due to the issues present in multiple-language processing, it made sense to focus on English for now.

Much like the extracted slur subsets, this project made use of multiple Common Crawl samples to produce a variety of results.

- Data sample 1 - First record found from all 60,000 segments
- Data sample 2 - First 60 segments, sampling records at a 25% rate
- Data sample 3 - First 200 segments, sampling records at a 20% rate

4.4 Common Crawl search

As we accessed the WET files, it was possible to extract the raw plaintext from the stream. This information was then decoded using the UTF-8 character encoding. This was selected as the decoding language of choice due to it having the highest percentage of use within websites, with most crawls containing more than 90% of all pages encoded using this character set, as stated by Common Crawl Statistics [10]. In addition to this, due to all of the slurs being converted to lowercase on extraction, it also makes sense to convert the plaintext to lowercase while also removing newline special characters and commas so it can be stored properly for the models.

Due to there only being one example for each of the slurs within the UD-20 dataset and the fact that most of these examples are of speech, which is a highly different form to what is present online, many models that require prior examples will struggle to determine whether a term is being used correctly or not. Instead, it made more sense to focus on preprocessing as an unsupervised sentiment analysis problem in a similar manner to Pamungkas, Basile, and Patti, 2022 [22]. From this, it was decided that using regular expressions to find text matches would be the most efficient approach before being analysed.

4.5 Sentiment analysis

As mentioned previously, in order to determine the use of offensive language within Common Crawl, a pair of unsupervised sentiment analysis tasks were completed.

4.5.1 K-Means Clustering

The first selected model was K-Means Clustering. This clustering algorithm could be used as a naïve method to cluster the found texts into either positive and negative, or abusive and non-abusive pages. This model can be adapted into additional clusters if a clear cut is not able to be established. In order to determine the clusters, this model makes use of the Word2Vec model from Gensim to automatically embed the sentences into lower-dimensional vectors. The produced vectors could then be used within the K-Means model; for this task, the model from scikit-learn was selected due to its high flexibility.

4.5.2 VADER Model

As an alternative to K-Means, VADER (Valence Aware Dictionary and sEntiment Reasoner) was selected to predict the sentiment of each of the found sentences. This model was selected due to the fact that it did not require a large amount of training data. This model is pre-trained and can handle commonly used acronyms and slang, which is common to find within internet blogs and pages. Unfortunately, due to the prior data processing, some features like punctuation and capitalisation have been lost, which will affect the model's predictive ability.

In addition to the model's predictive ability, the highly readable output for each sentence allows for a quick ranking system to be produced to compare whether certain terms tend to end up being used in more offensive or less offensive settings.

Chapter 5

Implementation

5.1 Neologisitic derogatory term extraction

5.1.1 Initial extraction and survey creation

As discussed in the previous chapter, regular expressions were used to collect the initial sample of slurs. The list of inclusive and exclusive terms is listed in Appendix B each of which was compared with the definition provided within UD-20. The inclusive matches were added, whilst the exclusive matches were removed from this subset to create the initial slur extraction.

From this, definitions with zero interaction values were removed, as if they were never interacted with on Urban Dictionary, it was unlikely that they would appear in the correct context online. The corresponding list was then run through a stratified sampling process, grouped by interaction values, before being sampled at a rate of 20%. From there, in order to get a small enough list for the survey, every 28th term was selected to produce 39 terms for the participants to look through.

5.1.2 Defining unusual terms

The results from the survey are shown in appendix C. From the vote distribution, it was possible to plot this against the interaction as shown in figure 5.1. As seen by the weak, positive correlation between the logarithm of votes and interaction, it is possible to utilise the votes to create a definition for unusual. The two subsets were determined based on these values using the following equations: where $\bar{v}$ is the mean of the votes, σ_v is the standard deviation of the votes, and v_n is the vote score of the n-th term.

$$\bar{v} - (\sigma_v/4) \leq v_n \leq \bar{v_i} - (\sigma_v/7) \tag{5.1}$$

$$\bar{v} - (\sigma_v/7) \leq v_n \leq \bar{v} \tag{5.2}$$

From these vote subsets, we can find the maximum and minimum interaction value shown below, where i_n is the interaction value for the n-th word. These ranges can then be applied to the full insult set to create two subsets.

$$1 \leq i_n \leq 22 \tag{5.3}$$

$$160 \leq i_n \leq 578 \tag{5.4}$$

5.1.3 Refining the list of slurs

Slurs consisting of less than 3 letters were removed, alongside definitions with a score lower than a quarter of their interaction value. In addition to this, terms with more than 3 words were removed due to the difficulty of implementing them within a model larger than a tri-gram. Finally, for the large-interaction dataset, dictionary terms were removed from the subset using the Enchant module, removing 18 terms. After all of these methods, the lower interaction subset consisted of 1100 terms, and the higher consisted of 101.

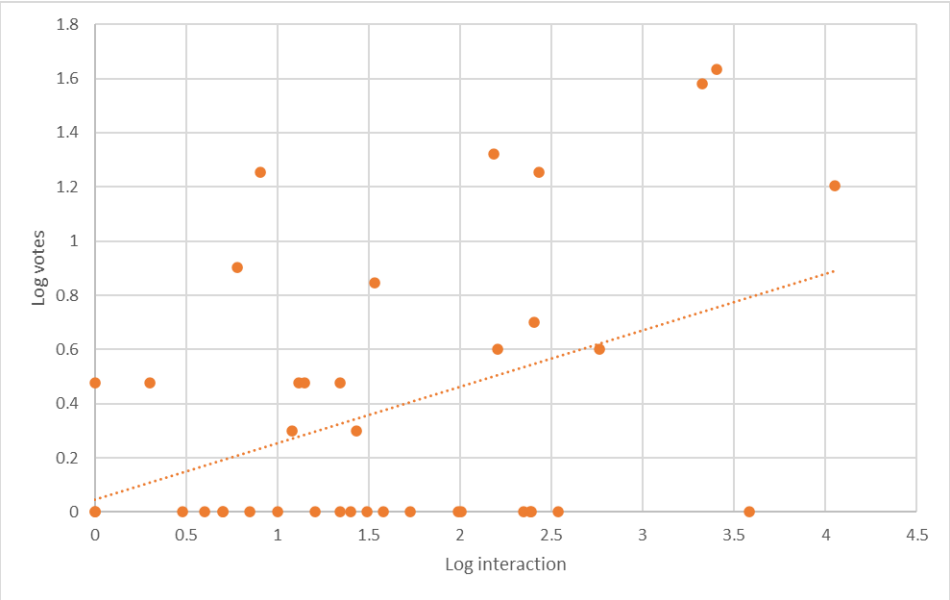


Figure 5.1: Survey results - log interaction against log votes by word

Parameter	Value
min_count	2
window	4
vector_size	300
sample	1e-5
alpha	0.03
min_alpha	0.0007
negative	20

Table 5.1: Word2Vec model parameters

5.2 Common Crawl subset and extraction

The index file for CC-MAIN-2020-29 was downloaded from the Common Crawl Index Server [9], allowing the segments to be downloaded through Python's requests module as a stream. Downloading the file as a stream cuts down on computation time significantly due to the decreased download effort as we only required the record headers and text of some of the smaller portions of the files.

To iterate through the downloaded files, a warcio archive iterator was used, accessing records when the header, "WARC-Identified-Content-Language", is set to English.

5.3 Sentiment analysis

5.3.1 K-Means Clustering

For the K-Means clustering task, the found texts were split into sentences, using full stops to separate them. These sentences were searched using regular expressions to find the sentences containing the insult, which were then formatted to replace non-alpha-numerical symbols with white space, whilst also reducing multiple consecutive white spaces with singular spaces.

These sentences are run through the Gensim Phrases function to detect common phrases, like multi-word expressions and word n-gram collocations. This output is fed into the Gensim Phraser which exports the Phrases model to a bigram model, to reduce memory usage.

From the bigram model, we can extract the formatted sentences to load into the Word2Vec model. This model converts the found sentences into lower-dimensional vectors that can be used within the clustering algorithm. The model was run with the parameters shown in the table 5.1.

The min_count tells the model to ignore terms with a frequency lower than the value, but due to the low sample size, the value is kept low for this implementation. The model is then trained to create the

Word	Sentiment Coefficient
a_roport	1.203629
the	20.659557
airport	4.053415
and	28.207516
of	17.953893

Table 5.2: Sentiment coefficient dictionary sample

vectors for the clustering model. The selected model was KMeans from scikit-learn which is capable of automatically separating the samples into their various clusters. For this task, the model was trained to cluster into two groups to logically separate the tasks into negative and positive sentiments.

Following on from producing the clusters, a sentiment coefficient value is calculated for each word in the vocabulary. This is completed by combining the cluster they are a part of and their closeness to the cluster. Table 5.2 is a sample of the terms and their respective coefficients.

Following this, the next step was to calculate a TF-IDF (term frequency-inverse document frequency) score for each word in each sentence. This was done to consider how unique each of the words is with respect to their sentiment. This was completed using Sci-kit learn's TfidfVectorizer.

By using both the TF-IDF scores and the weighted sentiment scores, we could replace each of the words within the sentences with these values to create two vectors. The dot product of these two vectors indicates the overall sentiment of each sentence.

5.3.2 VADER Model

Much like the beginning of the KMeans model, the found texts are split into subsequent sentences, using full stops to separate them. However, for this model, the previous and next sentences were also stored for the VADER model to make use of. The SentimentIntensityAnalyzer function from vaderSentiment was then used to analyse the polarity scores of the sentences.

After producing the negative, neutral, and positive sentiment scores for the various sentences. It's possible to average the values by word to determine which words tend to be used in a more negative or positive manner.

Chapter 6

Evaluation

This chapter covers the results produced by the various pieces of software and discusses the outcomes and some of their pitfalls.

6.1 Common Crawl slur matches

After searching the three CC datasets mentioned in the previous chapter, the total number of matches is shown in table 6.1, whilst the percentage of slurs that found a match is shown in table 6.3. Although we find the number of matches increasing massively between the datasets, the number of unique slurs increases disproportionately. Looking into the value counts of the matches, various are incredibly common within all three datasets due to their non-overt typical use; one such example is *Prosciutto* from the low interaction dataset, which appeared 356 times within dataset 3, as although it was defined as "a counterpose to gammon, a derogatory term used to describe followers of a hard brexit", most would consider prosciutto to be a thinly sliced Italian ham, and so many of the matches will not be representative. Similarly, from the high dataset, *Hearing Aid* matched 573 times within dataset 3, but these matches will more likely represent the typical definition rather than a "derogatory term for a bluetooth earpiece worn by anyone over 40 years old". These two instances demonstrate that there is a problem with the definitions of the derogatory terms. However, these instances will always exist since Urban Dictionary is an open-sourced website.

Dataset	Low matches	High matches
1	64	16
2	627	415
3	2487	1359

Table 6.1: Count of slur matches by dataset

Dataset	Percent low	Percent high
1	1.82	11.88
2	7.88	35.64
3	12.18	45.54

Table 6.2: Percentage of slurs with a match

6.2 KMeans Clustering

In order to test the clustering ability of this model, two metrics were selected to measure the cluster dispersion and the nearest-cluster distances. The Silhouette Score computes the mean intra-cluster distance and the mean nearest-cluster distance for each sentence; this value is best when tending towards 1, and is worse when tending towards -1, whilst 0 indicates overlapping clusters. The second metric was the Calinski and Harabasz Score [7] which is the ratio of the sum of between-cluster and within-cluster dispersion's. With this metric, a higher value is considered better, where there isn't an upper bound.

Slur dataset	Silhouette	Calinski and Harabasz
High	0.0088	1.3988
Low	0.0055	2.1213

Table 6.3: Common Crawl dataset 1

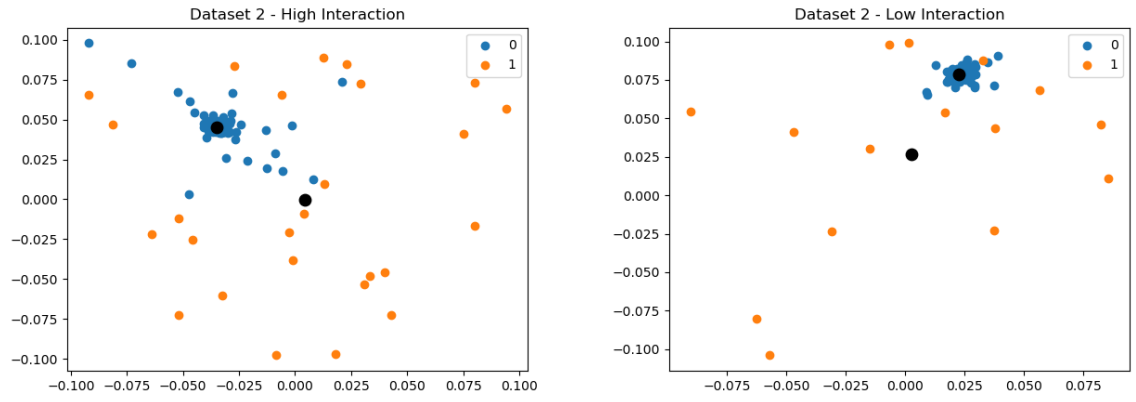
Slur dataset	Silhouette	Calinski and Harabasz
High	0.9776	5545.6
Low	0.9818	6673.2

Table 6.4: Common Crawl dataset 2

Slur dataset	Silhouette	Calinski and Harabasz
High	0.9802	8710.5
Low	0.9576	7951.1

Table 6.5: Common Crawl dataset 3

The three tables **??**, **??**, and **??** clearly show that dataset 1's clustering ability is significantly worse than datasets 2 and 3. This is most likely due to the dataset size, as it contains significantly fewer matches or examples. From the clustering within datasets 2 and 3, shown in figures 6.1 and 6.2, there is significant clustering of class 0, being the positive sentiment class, whilst the negative class has significantly larger and more distributed points. This initial insight demonstrates how a majority of the found sites tend to be positive or neutral in sentiment, even though this task was searching for terms which were defined as slurs. However, as mentioned in the prior section, many of the found terms would typically be considered non-insults, so this could be the explanation for this outcome.



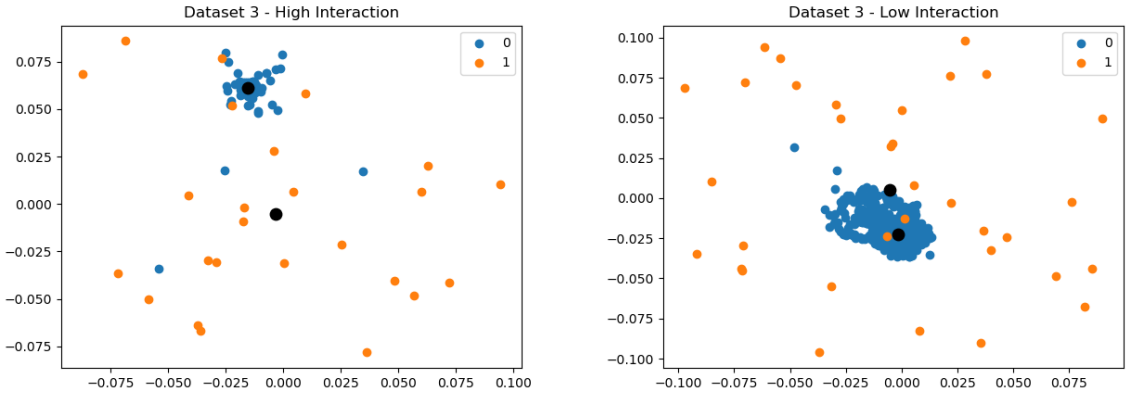
(a) High interaction clusters

(b) Low interaction clusters

Figure 6.1: Dataset 2 clusters

CC dataset	High interaction	Low interaction
1	11:18	65:54
2	1026:22	1015:37
3	2911:54	4960:155

Table 6.6: Positive to Negative sentiment ratios for each CC and Slur datasets



(a) High interaction clusters

(b) Low interaction clusters

Figure 6.2: Dataset 3 clusters

After calculating sentiment coefficient and TF-IDF scores, it was possible to predict the overall sentence sentiment by using the two values together. The table 6.6 presents the counts of the positive and negative sentences. As can be seen, much like figures 6.1 and 6.2, most of the predictions for the sentences within datasets 2 and 3 have positive sentiment, with only a handful of cases being negative.

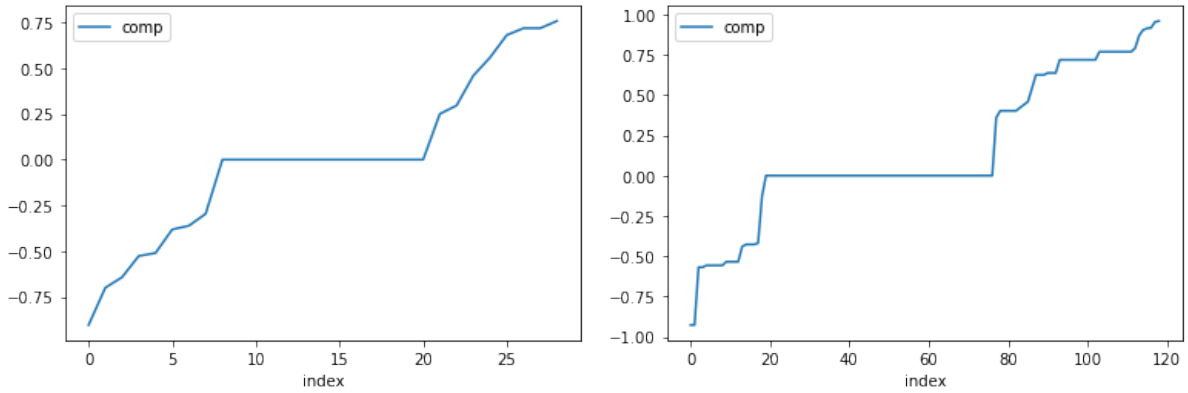
After taking a look at the cleaned sentences with the lowest sentiment score from dataset 3, many of the *worst* sentences were lists of terms. Two samples are: "suburbs aarons_pass abbotsbury abbotsford abercrombie abercrombie river aberdare aberdeen aberfoyle" and "joseph calvin college_cambridge college_cambridge jr_college camden county college cameron univ camp". Both of these texts are samples from a much larger collection of text that contains a number of terms that have negative sentiment scores. Pulling these samples together with the number of sentences predicted to have negative sentiment, we can conclude that this model has significant issues classifying sentences and lacks the accuracy needed to produce useful results.

6.3 VADER Sentiment Analysis

The VADER sentiment analysis model provides a negative, neutral, and positive sentiment score for each of the classified sentences, as well as a compound score that can be used as a single sentiment value. From this model, it was possible to highlight the overall sentiment trend within the found matches, as well as investigate the sentiment within individual terms and sentences. From figures 6.3, 6.4 and 6.5, it can be seen that the compound score of the three datasets tended to follow a similar distribution, with a significant portion of the scores being neutral and positive, much like the K-Means Clustering model.

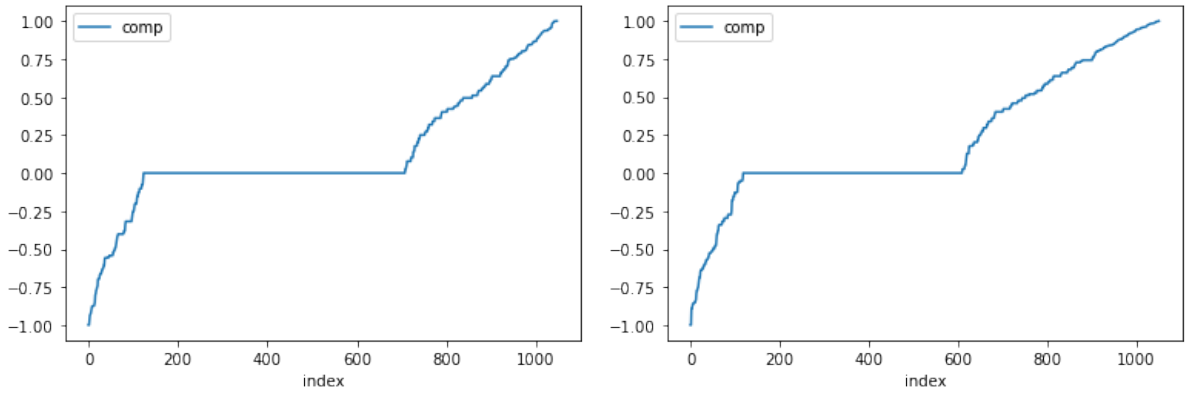
Investigating the terms and sentences with the lowest compound scores is an interesting way to delve deeper into the classified sentences. Focussing on the low interaction insults within dataset 3, the top 10 most *negative* sentences, according to the compound score, were based on websites containing pornography, video games , an archive of Welsh band names, and miscellaneous lists of terms which contain words like "firearms", "war" or "shooting" even though they are discussing matters like Olympic shooting events or games about war. From this small-scale sample, it is clear that this method of using VADER alone to classify the web pages will encounter issues without further formatting.

In addition to studying the individual sentences, it's possible to group the sentences by insult to aggregate the values, which shakes up the ordering of the most negatively scored terms. When we investigate the top ten *worst* terms, we find some overlap with some of the texts that contain pornography. However, many of the other texts discuss mundane and typical things like restaurant reviews and news



(a) High interaction (b) Low interaction

Figure 6.3: Dataset 1 ordered compound score by sentence



(a) High interaction (b) Low interaction

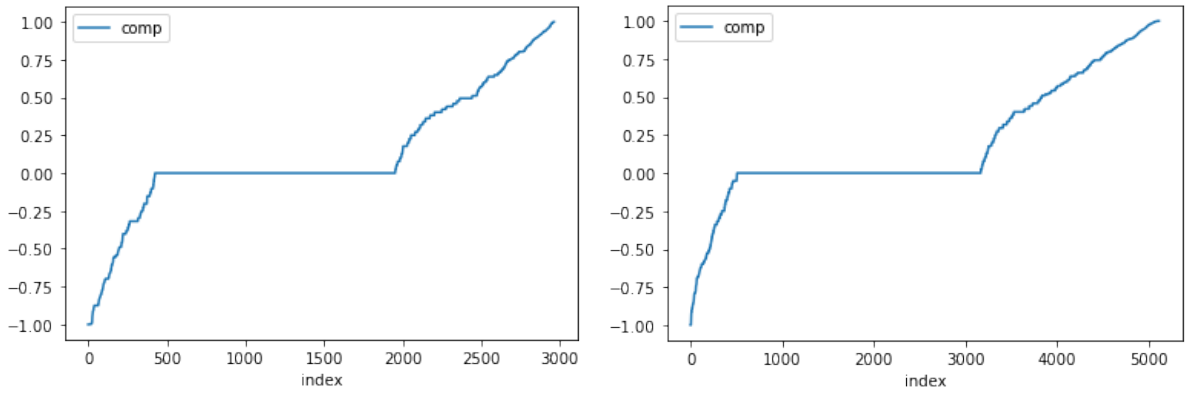
Figure 6.4: Dataset 2 ordered compound score by sentence

reviews. However, one of note contained the term "Saxon Dog". This term is the name of a blogger who is using this term in a reclematory sense as a form of identity, but it was highlighted as a negative sentiment due to their discussion of troops and war, but this is in the context of miniature tabletop wargaming, as it is definitely not being used in an insulting manner.

From the grouped terms, within the top ten most negative sentences, there was only one example of a term being used in a negative or insulting manner. The term being used is a concatenation of two traditionally insulting terms. With all of this in consideration, the low interaction slur subset has found minimal matches that could potentially be seen as a directed insult. However, this could be attributed to the neologistic nature of these terms or the extremely rare cases in which they are ever used in an insulting manner.

When analysing the results of the high interaction insults dataset, when aggregating the sentences together by term, unlike the lower interaction insults, many of the sentences they are used in are used in a derogatory sense. The high interaction subset is significantly more promising, ranging from forums discussing Special Operations to forums discussing comicbook art theft to religious forums discussing everything all at once. However, from table , the high percentage of terms with matches demonstrates the higher prevalence of these terms within the internet, with figure showing a larger *negative* tail within the graph, as these terms tend to be used in a more negative light due to their higher commonality.

Another section of analysis was completed on terms with only one match within dataset 3. From the high-interaction subset, there were six terms with only one match. After their sentences were analysed with VADER, two of those were found to be negative, two were neutral, and the last two were positive. However, on further inspection, all six of these appear to be negative in nature, with anti-semitism, harassment, typical swear-words, and leaked pornographic content discussions all being classified positively. This data further backs up the point that a more bespoke model trained on the definitions within Urban Dictionary may prove more fruitful in future work.



(a) High interaction

(b) Low interaction

Figure 6.5: Dataset 3 ordered compound score by sentence

6.4 Discussion

6.4.1 Extracting terms from Urban Dictionary

From the implementation shown in chapter 5, it is clear there are a number of different areas which could be improved upon or approached from different angles. One such area is the slur extraction from Urban Dictionary. The current approach of using regular expressions to extract insults from UD-20 is incredibly general-purpose, causing a great number of false matches within both the slur datasets. As an alternative, abusive language detection models may prove more fruitful, as discussed in the papers by Pamungkas *et al.*, 2022 [22] or Botelho *et al.*, 2021 [5]. However, many of the definitions and examples of neologistic slurs found within Urban Dictionary don't follow traditional syntax that is found within other data sources due to the extremely low user moderation found within terms with higher interaction and score. An example of this is the comparison between the definitions of "fuck" and "gewk":

Fuck: score: 50002
1. the universally recognized "f word"
2. n. implying complete and utter confusion
3. n. a really stupid person
4. v. to procreate 5. adj. can be used to modify any word for more passion
6. int. expresses disgust
7. int. expresses complete suprise and joy
8. adv. can be used to make a command more urgent

Gewk: score: 13
pronounced (gee-you-k)
gewk is in reference to the derogatory term for asians... gook.

From these two examples, it is very clear how Urban Dictionary's moderation works much better at scale, as the most up-voted definition for "Fuck" has multiple definitions organised into different linguistic categories, whilst "Gewk" is lacking in comparison. In addition to this, when comparing the highest scoring definition of "Fuck" with the lowest, which is shown below, the crowd-based voting system allows self-filtering to occur.

Fuck: score: -207
a beautiful word that should be used more often and accepted as the wonderful word that it is.

Another problem with using a crowd-sourced collection of definitions is the clash between what certain individuals define as a derogatory term and what others do not. At the time of release, UD-20 was moderated primarily by volunteers who decided what would and would not be posted. However, as mentioned within the Urban Dictionary Blog posts around this time, Urban Dictionary was suffering with moderation issues [23]. Along with this, there is an ongoing issue with language shift as people

attempt to define new insults and slurs that may have existed previously or would be considered non-insults to most. However, by ignoring these terms that have a *vanilla* use, this paper and task would ignore potential neologies and unique definitions, resulting in the model missing the initial task at hand.

6.4.2 Survey design issues

Following on from this, the survey encountered a few issues. As mentioned in the previous section, many of the terms defined as insults in Urban Dictionary also have a *vanilla* definition. This confusion between the insult and the typical definition caused issues within the survey, as participants anecdotally stated they had issues deciding whether they had heard of the term in an antagonistic sense or not. In future surveys, there are a number of different approaches that could be used to potentially abate this issue, including adding the term's definition as part of the survey to ensure participants are using the correct term definition. Further to this issue, the fixed slur sub-sample being used as part of the quiz only covered an extremely small portion of the overall slurs, and so developing a number of subsets or having a random sample could have covered a significantly larger portion of insults, and thus the bias towards known terms would be reduced.

In addition to the previous issue, there is also an issue pertaining to the participant pool size. As the participation was based on the recruitment advert and self-admission to the survey, it is likely that a self-selection bias was present in those that signed up, which, in addition to the fact that most participants were either students or young adults, the results of the survey will be askew. However, as the definition of "unusual" slurs is simply a placeholder for this study, as long as this is considered when studying the results, it should have minimal impact.

Another issue with the survey is the time at which it was done. As this paper is studying neologies or unusual definitions using a dataset for 2020, in order for the results to be more valid, the survey would have needed to have been completed at a similar time to the dataset's creation, as it's likely that use and definitions have changed over the years, and what would have been an unusual term back in 2020 may be much more prevalent in the current year. Keeping this in mind, the fact that the Common Crawl subsets that were examined were from the same time period as the Urban Dictionary crawl meant that some of this issue could be overlooked due to the age of insults used would still be accurate alongside the voting score.

With this in consideration of the survey's flaws, the alternative should be considered. If this paper hadn't made use of a survey, it may have been possible to utilise either the interaction or score values for the definitions to define a range for unusual terms. However, without direct user input from participants, we can't fully rely on the voting system within Urban Dictionary, as once again, due to its anonymity and the fact that it requires willing participants to vote on individual definitions, it potentially has biases within. However, these will still be prevalent in the current system as the project has made use of these scores all the same.

6.4.3 Refining the slur subsets

Following on from the survey, the lists of slurs were cut down in order to reduce the number of false-positive matches within Common Crawl, however, this had an impact on a number of legitimate insults that were extracted. The ratio between the interaction and score was chosen arbitrarily at four, however this could be selected as an entirely different value, or removed all together. Currently for the high interaction set, this removes 42 out of 165 terms, whilst for the lower set, 1011 out of 2133 are removed, which is a very significant portion. An issue of contention is within the trimming of the high interaction subset; terms which appear in the dictionary were removed due to the their high prevalence of false-positives, even though they could legitimately be considered slurs by some, but due to the sheer number of false positives within the Common Crawl search, they were removed all the same. Some examples of these terms were: "Covid - A derogatory term for a child conceived during the Covid-19 pandemic." and "Cant - As in the derogatory term for the female genitals. Used as an insult. The 'A' can sometimes be lengthened in sound." Both of these examples would have been deemed legitimate within the dataset due to their score of 314 and 62 respectively.

6.4.4 Searching Common Crawl

When searching Common Crawl, much like the slur extraction, a naïve approach was selected by simply using regular expressions. In a similar vein to the slur extraction, this method could be completed using different methods that either use a modified sentiment analysis model pre-trained on other hate

speech detection tasks, with the extracted slurs added into the model with pre-defined negative sentiment scores; alternatively, it could be possible to make use of the definitions and examples from within Urban Dictionary within a separate model, but due to their conversational-based and un-moderated crowd-sourced definitions, this approach could prove more of a hinderance.

As an entirely separate approach, it could be interesting to instead approach this problem from well-known websites that contain hate speech. Similar to the work by Bogetić, 2022 [4], starting with the websites of known hate sites, like the now-defunct Kiwi Farms or Stormfront, it may be possible to trawl their websites for links to accompanying or similar pages in search of smaller communities. Although, this could prove promising by starting from forums that are already known to contain hate speech, it may still prove difficult to search for neologies due to their uncommon presence.

Chapter 7

Conclusion

This chapter summarises the work covered within this paper, re-affirming the project's aims, discussing the current status of development, the outcomes and future work that could be explored.

7.1 Content review

For this project there were three primary aims:

1. To define what could be considered an *unusual* derogatory term
2. To locate websites or forums of note that are using these terms
3. To determine whether a term was being used an an insult or in a reclamatory sense

7.2 Project status

Regarding the first project goal, it was possible to define a pair of interaction boundaries to define two sets of derogatory terms using the voter interaction score from within UD-20 and votes from the participant survey. From the sentiment analysis, it's still unclear as to where a proper definition should be laid, but for this paper, the contrasting ranges allow for an interesting insight into the sentiment analysis models.

For the second aim, it is possible to say that it has been partially completed. Within the Common Crawl search, a small number of pages were found to be using the *unusual* insults from the higher-interaction subset within dataset 3 through the means of the VADER sentiment analysis model. One such example comes from http://www.socnet.com/ where a user by the name of "five-o" used the term "meat gazer" in an insulting manner. Although a minor result, the fact that the model was able to identify a page using this term correctly is very promising. However, there are still issues with the classification, as although this was correctly identified, there were sentences which were entirely inane but were still flagged as negative due to singular words like "devil" or "war." In addition to this, although the model was able to identify negative sentiment, the final analysis required a large amount of human-based interaction in order to find specific pages, but this could be potentially automated with a bespoke model or additional modelling.

For the final aim, the models were unfortunately unable to distinguish between negative insults and those being used in a reclamatory manner. However, through individual analysis, it was possible to find a small number of cases where insults were being turned around and used in an ironic manner by those that would have been targeted previously. One such example is Saxon Dog, a blogger whose name is considered an insult by Urban Dictionary, but it was found that the term was instead being used in jest.

7.3 Further Work

When looking towards future endeavours, there are a number of different tasks and approaches that should be considered for future development. One idea mentioned earlier in the project's aims is to trawl through Common Crawl, collecting URLs from pages with negative sentiment and then studying the page links within to build a network and hopefully determine how certain terms spread within communities. Alongside this, in order to improve the search, a study on the effect of altering the filtering of the slur

subsets, with singular word slurs or combinatorial terms being analysed on their own to see the effect on the number of matches found, could prove useful in determining the amount of filtering that should be applied to the subset created from Urban Dictionary.

Alongside this, work on the sentiment analysis model could prove fruitful in creating an informal language detection model, but there may be better sources for gathering conversational or slang-based text.

Bibliography

[1] Sai Aluru, Binny Mathew, Punyajoy Saha, and Animesh Mukherjee. *Deep Learning Models for Multilingual Hate Speech Detection*. 2020.

[2] The World Bank. Individuals using the internet (

[3] Anna Belkova. Neologisms formation using borrowed affixes in the russian internet segment. *SHS Web of Conferences*, 55, 2018.

[4] Ksenija Bogetić. Race and the language of incels: Figurative neologisms in an emerging english cryptolect. *English Today*, pages 1–11, 2022.

[5] Austin Botelho, Bertram Vidgen, and Scott Hale. *Deciphering Implicit Hate: Evaluating Automated Detection Algorithms for Multimodal Hate*. 2021.

[6] Luke Breitfeller, Emily Ahn, Aldrian Obaja Muis, David Jurgens, and Yulia Tsvetkov. Finding microaggressions in the wild: A case for locating elusive phenomena in social media posts. In *EMNLP*, 2019.

[7] T. Caliński and J. Harabasz. A dendrite method for cluster analysis. *Communications in Statistics*, 3(1):1–27, 1974.

[8] Bianca Cepollaro and Dan López de Sa. Who reclaims slurs? *Pacific Philosophical Quarterly*, 2022.

[9] Common Crawl. Common crawl index server, 2022.

[10] Common Crawl. Statistics of common crawl monthly archives, 2022.

[11] Adam M. Croom. How to do things with slurs: Studies in the way of derogatory words. *Language Communication*, 33(3):177–204, 2013.

[12] Andrew Dai and Quoc Le. Semi-supervised sequence learning. 2015.

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota, 2019. Association for Computational Linguistics.

[14] Kevin Heffernan, Onur Çelebi, and Holger Schwenk. Bitext mining using distilled sentence representations for low-resource languages. *arXiv preprint arXiv:2205.12654*, 2022.

[15] Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 328–339, Melbourne, Australia, 2018. Association for Computational Linguistics.

[16] Ian Johnston. *Neologisms and their use in gaming communities*. PhD thesis, Southern Illinois University Carbondale, 2021.

[17] W. Liu and Wenyu Liu. Analysis on the word-formation of english netspeak neologism. *Journal of Arts and Humanities*, 3:22–30, 2014.

[18] Alexandra Luccioni and Joseph Viviano. What's in the box? an analysis of undesirable content in the common crawl corpus. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), pages 182–189, Online, 2021. Association for Computational Linguistics.

[19] José Mateo Martínez and Francisco Yus. Towards a cross-cultural pragmatic taxonomy of insults. 2013.

[20] Colin Morris. Compound pejoratives on reddit – from buttface to wankpuffin, 2022.

[21] Dong Nguyen, Barbara McGillivray, and Taha Yasseri. Emo, love, and god: Making sense of urban dictionary, a crowd-sourced online dictionary. *Royal Society Open Science*, 5, 2018.

[22] Endang Wahyu Pamungkas, Valerio Basile, and Viviana Patti. Investigating the role of swear words in abusive language detection tasks. *Language Resources and Evaluation*, 2022.

[23] Aaron Peckham. Update of content moderation, 2021.

[24] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 2227–2237, New Orleans, Louisiana, 2018. Association for Computational Linguistics.

[25] Alec Radford and Karthik Narasimhan. Improving language understanding by generative pre-training. 2018.

[26] Steven Wilson, Walid Magdy, Barbara McGillivray, Kiran Garimella, and Gareth Tyson. Urban dictionary embeddings for slang nlp applications. Proceedings of the 12th Language Resources and Evaluation Conference, pages 4764–4773, Marseille, France, 2020. European Language Resources Association.

[27] Steven R. Wilson, Walid Magdy, Barbara McGillivray, and Gareth Tyson. Analyzing temporal relationships between trending terms on twitter and urban dictionary activity, 2020.

Appendix A

Github repository and video link

Repository location: https://github.com/ItsmeHJB/DataScienceProject

Video link: https://drive.google.com/file/d/1ADQqVmqvutSMr8AZWKucCIqz7nHcube1/view?usp=sharing

Alternate video link: https://youtu.be/6NnznOr9YjA

Appendix B

Inclusive and Exclusive extraction phrases

Inclusive phrases	Exclusive phrases
derogatory term	shutting down racist
insulting term	useful term
racist term	to slur your
prejudice term	to slur their
prejudiced term	to slur his
a slur	insert derogatory
offensive term	always slur
	slur over
	slur pee
	you slur
	intentionally slur
	from a slur
	friendly and tender
	strong feeling
	non racist term
	non derogatory term
	not a slur
	not a racist term
	was a derogatory term
	with affection
	from the common derogatory term
	not a derogatory term

Table B.1: Inclusive and Exclusive phrases used in term extraction

Appendix C

Survey results

word	hippapottafrog	mr. fist bump	moldy tip	jj redneck	tose me	flipcracker	dingly wert	branch weight	microballs	applethrower	nancy no-tits	hekma	chewgger	ice gook	robuttnik	njf	blanger	man bum thumper	poputt
interaction	1	1	1	2	3	4	5	5	6	7	8	10	12	13	14	16	22	22	25
vote	0	1	3	3	0	0	0	0	8	0	18	0	2	3	3	0	0	3	0

gewk	chundee	slutbasket	kopitard	gickna	fishcake	pstool	curtain twitcher	touch down	pepperbelly	cape	gunga	collar popper	schoolgirl	fbs (fat bitch syndrome)	chit	oriental	ass-hat	squid	g
27	31	34	38	53	97	101	152	160	221	241	246	255	269	345	578	2114	2530	3829	11256
2	1	7	0	0	0	0	21	4	0	1	0	5	18	0	4	38	43	1	16