ILLICIT EUPHEMISMS ONLINE

How can natural language processing methods be integrated into efforts to moderate euphemistic illicit speech online?

QKVJ3 – BASC0024

May 2022 University College London | Supervised by Prof Lewis Griffin

Abstract

Euphemisms are a way for fringe and criminal entities to escape common methods of content moderation in online spaces. This paper investigates how these euphemisms could be better detected in online moderation with automation. First and foremost, the problem area and the different considerations surrounding it are defined. The definition of illicit speech, euphemisms, as well as the motivations of different groups participating in illicit speech are defined. The problems faced in content moderation are discussed, and this paper asserts that there is a need for more automated moderation. This would be done using natural language processing. Using 'Self-Supervised Euphemism Detection and Identification for Content Moderation' by Zhu et al. as a test case, the functioning of a euphemism detection solution is experimentally evaluated and used to further the conversation on how euphemism detection would be implemented in practice, also considering the intricacies of multi-format and multilingual communications. Finally, a novel content moderation loop framework is proposed and discussed, asserting that an automated content moderation framework must consider all the elements discussed in this paper in order to be viable.

Preface

This paper aims to investigate the use of euphemisms – veiled or codified speech intended to be understandable only for some people – in online communication spaces. This phenomenon is especially pertinent in large online spaces where multiple communities collide, and thus there is both unintentional and intentional instances where language becomes in-group specific. A harmless example of this would be fandoms, that often develop language specific to their interest that outsiders may not understand.

This paper, however, does not consider such innocuous uses of euphemisms, as there is a growing need to regulate speech with malicious or criminal intent. With the rise of the alt-right and other hateful movements such as incels, it is more important than ever to keep track of what interactions take place in large online forums. A recent example of a euphemism with malicious meaning is Pepe the Frog. Originally just a cartoon, in the past two decades it has been reappropriated by members of the alt right to create hateful memes, and in a larger sense, to symbolise their movement. The danger that lies in this style of covert communication, is that not only is it easy to deny any serious intent and claim satire, but it also serves as a gentle introduction to extremist ideologies. With edgy humour and the ridicule of outsiders as part the first steps in alt-right radicalisation, having veiled extremist communications on large forums poses a risk for larger public safety.

Recent developments in natural language processing enable researchers to both produce and analyse natural language more accurately than ever. In the context of online content moderation, where human moderators struggle to keep up with the large amounts of content published every day, natural language processing may provide solutions that can handle content a lot faster and in an automated way.

The use of natural language processing for detecting euphemisms is a growing field with many ground-breaking papers being published in the last few years alone. One of them, Self-Supervised Euphemism Detection and Identification for Content Moderation, is used as an example of the potential that this research field has.

Although the solutions discussed derive from computer science, this paper is inherently interdisciplinary with context being drawn from internet policy, law, linguistics and sociology. On a personal level, the paper draws on knowledge I have acquired on the Arts and Sciences degree in computer science, language studies, as well as internship research into the dynamics of extremist movements.

Table of Contents

1. Background and context

1.1 Scope of illicit speech as defined by platforms

Online platforms facilitate interactions between people, which inevitably leads to conflict over what interactions are permitted. Certain content is deemed illicit; notably, content that platforms and regulatory bodies are interested in moderating, and which potentially causes harm to other civilians. It is important to clearly define which content is deemed illicit so that it can be best acted upon.

As one of the most prominent platforms for online discourse, Twitter's (Twitter Help Center, 2019) code of conduct specifies that threats or promotion of violence, violent extremism, child sexual exploitation, targeted abuse and harassment, hateful conduct on the basis of protected characteristics, suicidal, graphically violent, or pornographic content, and the distribution of illegal goods and services, are forbidden. Moreover, the spread of misinformation through "platform manipulation and spam" and "synthetic and manipulated media", or for the purpose of interfering with elections, is also regulated. Twitter's definition of illicit speech will serve as the framework for this dissertation's discussion of it.

In the context of misinformation, the Centre for Data Ethics and Innovation (2021) distinguishes between deliberate spreading of false information (disinformation), and inadvertent spread of false information. While the former implies a more worrying trend, identifying which of the two is the case often requires investigation. In general, they will at first hand be dealt with in the same way - with misinformation flairs and removal in more severe cases.

The scope of what constitutes illicit speech is thus extremely large and varied. With further nuance within each category on what exactly fits the definition, the task of moderating such content is complicated, and many of the sources mentioned in this paper will focus on a certain type of illicit speech. For the purposes of this dissertation, it will be considered that while the types of speech vary greatly, the end goal is the same - detecting and removing the content where necessary. If a certain moderation case only pertains to a certain type of illicit speech, it will be explicitly stated. Defining each subtype of illicit speech in detail would be challenging, which is why this dissertation will use external definitions of each type on a case-by-case basis.

1.2 Motivations

Understanding the motivations of those that engage in illicit speech is crucial in defining the problem scope. Not all instances of illicit speech are malicious, however, there are several malicious communities that engage in illicit speech online, with overlaps between groups and types of illicit speech.

For hate speech (what Twitter would define as threats or promotion of violence, targeted abuse and harassment, or hateful conduct based on protected characteristics), a strong motivation to confirm in-group mentalities and bond over shared ideology (in this case antagonistic views towards another group) may explain the persistence of it.

Putnam (2000, p. 21) writes that network and social reciprocity are extremely beneficial for the in-group, however, they can also be identified as the phenomenon that causes the creation of hateful in-groups such as the Ku Klux Klan, as the same principle applies for any group that is formed on the basis of shared motivations. Putnam further states (p. 23) that bonding social capital, while effective in creating in-group loyalty, may as a result also create strong out-group antagonism. Speaking of internet chat groups, Putnam states that the in-groups will often be tightly homogeneous in education or ideology, rather than other personal characteristics. Speckard et al (2021) found that members of the incel movement (involuntary celibates), were more likely to have their misogynistic identity strengthened in proportion to the frequency of their participation on incel forums, where they were able to share their views with other incels. These communities enable the violent tendencies of some incels, and previous incel perpetrators of violence such as Elliot Rodgers were shared on these forums as inspiration.

For misinformation, Hindman and Barash (2018) report that although there are numerous small-scale actors who publish fake news with the aims of gaining some ad revenue, these fake news are not the main factors in the misinformation ecosystem. At the centre are political entities that want to influence the outcome of elections and play a part in world events. Hindman and Barash state that there is sufficient evidence to indicate that the 2016 U.S. elections were influenced by the Russian government through tens of thousands of social media accounts sharing fake stories, thousands of professional trolls and digital ads targeted towards American users - as well as 381,369 news stories, according to Hindman and Barash's calculations (p.14). They also note that such large-scale efforts to confound public opinion have been observed in other countries also, indicating that there is widespread motivation to spread misinformation.

For the sale of illegal goods and services, large private platforms provide access to a large demographic to market to, including young teenagers and adults, who are the main buyers for illegal drugs (McCulloch and Furlong, 2019).

According to the United Nations Office on Drugs and Crime and United Nations Counter-Terrorism Implementation Task Force (2012), terrorist organisations use online recruitment to target marginalised members of society, especially minors. This would entail infiltrating online spaces where the target demographic is active and blending the propaganda between other, innocuous communication.

Danskin (2019, 1:20) details how the alt right in the age of the internet is an extremely decentralised movement, with no single figurehead, and no single community where they inhabit. This decentralised nature not only multiplies the reach of the ideology but makes it more challenging to track and define. Danskin's 5-step radicalisation framework describes the alt right's process in finding communities where their target demographic - most often white men who feel isolated from society - are present and infiltrating those communities with rhetoric that ridicules and others leftist ideology and the demographics that the radicalisation targets are starting to build resentment towards. Here is where the social reciprocity of hate speech plays a part, strengthening the target's belonging in the alt right circle. The more of the content the subject consumes, the more normalised it will become.

The platforms own recommendation systems are part of the problem; they have been known to recommend to users' content that according to their own rules should have been removed. This can lead unsuspecting users down rabbit holes of increasingly radical content that they would seek themselves. Mozilla Foundation (2021) found in their study that out of the 3362 "regrettable" YouTube videos flagged by study participants, 71% were recommended by YouTube's algorithm. With conspiracy videos, sexualised videos masked as children's content, and promotion of white supremacy, the videos represented the kinds of illicit content that violates YouTube's community guidelines, as determined by Mozilla. Only 200, or about 9%, of the flagged videos ended up being removed by YouTube. Mozilla highlights that the total view count of these 200 videos was over 160 million prior to deletion, meaning that a large audience had already been impacted.

A crucial aspect that Danskin discusses is that the radicalisation process works in layers. Initially, the content viewed by the soon-to-be-radicalised user is assumed to be satirical, or "trolling", with not much thought given to the possible deeper meaning behind it. With the repetitive nature of internet memes and media, the target will become desensitised to the more shocking jokes, thus moving further down the radicalisation pipeline. Analysing any instance of a right-wing meme or media piece, it would be difficult to tell whether the poster's intent is satirical or not. This is another key element to the alt right's functioning: someone who is not particularly extremist, but rather on the edges of their ideology, may inadvertedly play into their hands by sharing a meme or using a turn of phrase popularised by the alt right. This process could be described as "self-radicalisation" – other than what can be attributes to the recommendation system, the radicalisation target seeks out the content themselves, only consumes what they are in the moment comfortable with, and willingly participates in the discourse. The dangers of extremist communities overlapping with mainstream internet communities are thus apparent: if the circumstances align, anyone is at risk of falling victim to radicalisation.

Although the groups that deliberately engage in illicit speech on online platforms all have different aims, their shared motivation is to evade moderation.

1.3 A workaround: euphemistic speech

For all these distinct groups and types of illicit speech, the desire to evade moderation leads to alternative ways of communicating their intent. Veiled language will be used as a way to remain undetected. The term used in this paper for this kind of speech will be **euphemisms**, or veiled speech. The term describes a kind of speech that is deliberately vague. It can be broadly described as language that looks to mean one thing to an uninformed reader, usually something innocuous, and another to an informed reader.

A prominent example of euphemisms is the use of political dogwhistles. Stanley (2015, cited in Saul, 2018) presents that covert dogwhistles are often introduced as common ground in conversation in a way where the receiving party may not be aware or able to articulate their opposition. Mendelberg (2001, cited in Saul, 2018) researched how after the 1960s straightforward racist messaging was no longer acceptable, as voters did not want to think of themselves as racist - although the same biases were still present. Hence, the use of subtle and indirect messaging that still covertly carried the racist intent was more effective. On the other hand, an overt and intentional dogwhistle is specifically designed to mean something different to in-group and out-group listeners - although these are less in use in current politics. Former U.S. President Donald Trump is especially known for his use of dogwhistles to appeal to voters. Burack (2020) raises how terms such as "international bankers", "global financial powers" and "special interests" were all used in Trump's speeches as antisemitic dogwhistles - these are overt and intentional, because his words are expressly targeted towards antisemitic voters that recognise the rhetoric.

As detailed in Zhu et al. (2021), euphemisms are widely used in selling illegal goods and services online. For example, words denoting drugs keep changing, as previous euphemisms become known and get banned. "Ice" can mean methamphetamine, but so can "chalk", "crank", or "wash".

Bhat and Klein (2020) present that on Twitter, right-wing groups have certain methods of dogwhistling - or euphemistic speech - that they are able to use to circumvent moderation and reach their goal of signaling community belonging and reach a wider audience with their ideologies. They use both historical right-wing symbolism as well as common media symbols to adhere to Twitter's code of conduct. By using catchy imagery and phrases they are able to appeal to an audience that would not accept the ideology at face value. This is how the radicalisation process described by Danskin (2019) employs euphemistic speech - it serves a double purpose of signalling to other in-group members present, but also serves as a soft introduction to the group for potential recruits.

The dangerous nature of euphemisms is that they would not be illicit if it were not for the context around it. The word itself, such as "ice", is not what makes the euphemism, but the surrounding sentence or discussion ("This ice will get you so high") makes the real meaning apparent for an informed reader. In some cases, an entire sentence needs to be read in the context of the surrounding discussion or the identity of the poster for the meaning to become clear. This makes moderating euphemistic illicit speech a task that needs to be clearly defined in order to get accurate results.

2. The dilemma of content moderation

The issue of what content needs to be removed and what does not, is more nuanced than simply categorising types of illicit content to be acted upon. Twitter Help Center (2019) assures that they are considering the "larger story" surrounding a Tweet when deciding on its removal. This, however, also means that the enforcement process is going to be slower and more expensive. This specification also implies that the definition of what kind of speech belongs in these categories can be quite vague, and at times arbitrary. Gillespie (2018) recommends that moderation guidelines should be viewed as "provisional lines drawn in shifting sands", as they are but ever-evolving compromises.

Gillespie (2018, p.75) raises how little data there is on the true scale of platform moderation. It is unclear how much involvement user reporting has in the flagging process, and how much of the moderation work is done manually. Gillespie recounts that most platforms favour an approach where all content is allowed to be published and will only be moderated once it is live. This means that even the most illegal and hateful of content will appear on a site before anything is done about it. This level of unclarity in moderation processes makes it difficult to evaluate from the outside where exactly the weaknesses of current moderation systems lie.

Another side of the moderation debate is the necessary regulation concerning illicit and illegal content. Governments must establish oversight mechanisms in order to ensure the responsible governance of online speech.

In the case of misinformation, platforms are not legally required to remove it, unlike illegal content such as selling drugs. This means that they have more freedom over how exactly they choose to deal with it. Content that is "offensive, objectively wrong, or even carry some risk of harm" (Centre for Data Ethics and Innovation, 2021) can still be allowed on a platforms if there is no clear policy surrounding its removal. Large platforms especially can be reluctant to remove or over-police content as this would also be harmful for their brand image. Thus, content that is not clear-cut in how harmful it is, such as euphemisms that would be more challenging to prove as harmful, is less likely to be removed. There is a challenge in determining what the appropriate balance is between removing misinformation and seeming too authoritarian.

Although platforms have major freedom in how they regulate content, that is not necessarily how it should be. Land (2019) asserts that privatising censorship does not lead to harmful content being removed as effectively as it should be. Land presents that although governments have clear laws on what constitutes illegal content, the responsibility to review and remove such content is on the platforms themselves. The European Union as well as multiple national governments in the EU such as Germany and France, have laws which state that to avoid liability a platform has an obligation to monitor for clearly unlawful content (as mandated by the European Court of Human Rights), and to reactively remove flagged content (as detailed in the EU e-Commerce Directive). Land (p. 396) argues that this is not enough and can easily lead to human rights violations under international law. Governments have a legal obligation under international law to both safeguard its citizens and protect free speech, but private companies do not have the same obligation, and thus are more likely to moderate content in a way that suits their goals. With the government-mandated responsibility of moderation falling on private platforms, there is no real guarantee that they would completely adhere to what can legally be considered the government's responsibility to safeguard its citizens.

Another large question in content moderation lies in the global nature of the internet. With no state borders, platforms are responsible for adhering to the regulations of the countries they are available in. Therefore, it is especially pertinent to consider the implications of international law being applied to the regulation of online speech.

Land also raises (p. 407) that there is no such thing as perfect enforcement of wrongful speech, neither in an offline - nor online context, as speech regulation treads a fine line with respecting freedom of speech. Unfortunately, this means that in some aspects questions of content moderation will not reach an airtight solution - and euphemistic speech is extremely likely to evade moderation as it is more difficult to prove that it is indeed in violation of platform regulations or law. This is also why there should be greater transparency from platforms in what exactly their content moderation processes entail, as that would provide some accountability for clearly removing illicit content.

NYU claims (Centre for Data Ethics and Innovation, 2021, p.21) that there should be more than double the number of human moderators as now - but they should also not be outsourced and be given genuine worker's benefits. According to the Centre for Data Ethics and Innovation

(2021), TikTok, Facebook and Youtube have issued statements that the current increased reliance on algorithms that can be observed is only temporary, and that human moderators will continue to be at the core of their processes, as they are not interested in implementing a more algorithmic approach. This preference for human moderators is expensive and has a cost on the people doing the work. As seen in Cradle (2021), the tremendous mental health impact that moderation has on individuals has resulted in legal repercussions for Facebook, that had to pay compensation for outsourced moderators that suffered from severe PTSD for the violent imagery that they viewed on the platform, with very little support offered from Facebook to mitigate this.

Due to this human cost and concerns for scale, accuracy, and legal uncertainty, this paper asserts that a more algorithmic content moderation style is necessary.

The Centre for Data Ethics and Innovation (2021, p.18) determines three distinct uses of algorithms in content moderation. In filtering, content is not allowed to go live if it is deemed to be in a banned category using AI checks. This generally means that there is a delay in content going live, which may make it less popular with platforms that generally want to prioritise user experience. In detecting, content is flagged and prioritised to be later reviewed by a human moderator. In removal, AI is given the power to decide if content should be removed or not.

In Zhu et al. (2021), they state that Google's automated toxicity detector, Perspective API, is used by content moderators to prioritise content to be reviewed based on toxicity scores. In Zhu et al. 's evaluation of the efficacy of their euphemism detection model, they used Google's Perspective API to check the toxicity scores of both non-euphemistic drug selling, and euphemistic drug selling. Their results in this experiment suggest that if a moderator is using Perspective API to filter and prioritise content, they would be less likely to see the euphemistic content, as it generally had lower toxicity scores. This indicates a gap in the efficacy of automated systems: they may be designed to handle content where the illicit nature is unambiguous, but content that is designed to confound even human readers adds another layer of complexity in detection.

With such large scales of content being published at all times - in 2020, Twitter saw an average of 500 million tweets being published per day (Sayce, 2020) - it is inevitable that "fringe" cases will still be represented by a large amount of content. There is both a risk in this content remaining undetected and being viewed by other users - and a potential benefit to automated moderators that would be able to apply standardised takedown decisions to repetitive illicit content. The large amount of content published online every day calls for faster moderation methods. Therefore, it is crucial that the potential of machine-based solutions is understood in the scope of this problem. The potential solution for content moderation lies in natural language processing.

3. Natural language processing

The problem with natural language processing is how language can be understood by a machine to a high degree of accuracy. Generally, they take a text input that they then ‘encode’ into vectors, and work with the numerical data. It is only at the end where the ouput is ‘decoded’ back into words. Working in this encoded multidimensional vector space allows for efficient analysis. Until recently, the common framework for language related tasks was a recurrent neural network (RNN), with an encoder-decoder architecture that was sometimes connected to an attention mechanism.

The encoder-decoder architecture takes some input, such as a sentence, and encodes it into an embedded vector format, which is what the model reads. The decoder on the other hand translates the embeddings back into an output that a human can read (Saeed, 2021).

Bahdanau et al. (2014) proposed the attention mechanism to improve the efficiency of the encoder-decoder. It takes a weighted sum of all the hidden encoder states to inform the decoder what the most relevant information is.

An RNN, however, is extremely slow to train. It can only work with sequences in relation to each other - it needs information on the previous state to operate on the current state.

Because of this, the gradient descent calculations that ensure nodes with higher error rates are given less weight become less accurate. More specifically, this means that the gradients start to vanish or explode the further back the backpropagation goes in the neural network. This leads to loss of information from earlier sequences. In the context of natural language processing, this might mean that if a word in a sequence refers to something in earlier states it may not make the connection. For example, if in *I saw the dog today* the dog is referring to a specific dog introduced somewhere earlier, the RNN may not be able to retain this connection, as it is only able to move over the sequence one word at a time.

This changed, however, when Vaswani et al. (2017) proposed the transformer for language translation. The transformer is a simpler architecture that only makes use of attention mechanisms. In language processing, where RNN needed to pass each word at a time, the transformer is able to take an entire sentence at a time and return all the word embeddings at once. This will also include positional information, as the same word in different positions may mean different things.

The transformer approach solved the issues surrounding the former methods - it is faster as it is able to accept larger sequences as parallel input, and it is bidirectional, meaning that it can learn context from both past and future states i.e. a sentence is evaluated in the context of both the preceding and following sentence.

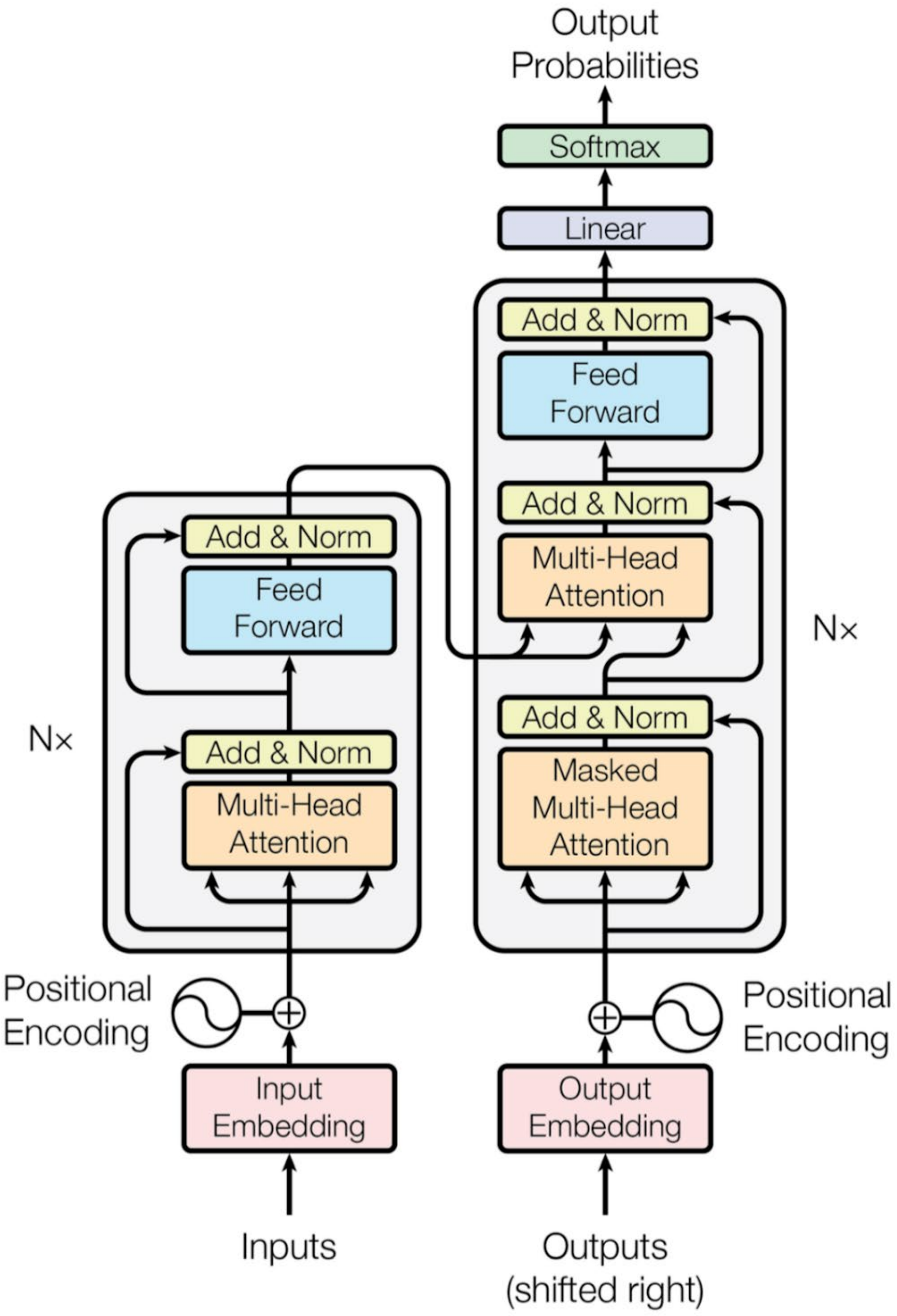

Figure 1: The Transformer - model architecture.

Figure 1: The transformer model architecture (Vaswani et al., 2017)

An important element of the transformer is that the encoder and decoder tasks are distinct from each other. Although they work together, both are equally powerful in their respective tasks: through training, the encoder learns grammar and context, whereas the decoder learns to map out connections between the input and output, such as mapping word connections between two languages in a translation task.

It is this distinction that allowed the creation of the Bidirectional Encoder Representation of Transformers, or BERT, which as the name suggests, is a stack of transformer encoders. First proposed by Devlin et al. (2018), BERT is capable of much more than just language translation. By pretraining BERT to understand language and then fine tuning it on a downstream task, one can solve many language-related problems.

3.1 Euphemism detection for content moderation

One such downstream use of BERT is by Zhu et al. (2021), who have pretrained a BERT model to detect and identify the meaning of euphemisms based on context. Their proposed approach would be used to assist in the creation and maintenance of ban lists - words that are automatically picked up by a moderation system and banned. They propose the creation of a "content moderation pipeline", which takes raw text and target keywords as input, and the detection and identification outputs will be used by moderators in their work.

The salience of their approach comes from its originality in tackling the problem. They pretrained BERT on a large corpus of Reddit posts that contain euphemistic speech, which prepared the model for handling unlabelled data. At the detection stage, the target keywords in the input sentences are masked out, as the model works with the contextual information surrounding the mask. Euphemism detection and identification are difficult tasks because of the effort to distinguish between innocuous uses of the word and genuine euphemisms. This difficulty is addressed in their research by using a Masked Language Model (MLM) to filter out sentences where the context is too generic to accurately study the euphemism. The MLM finds candidate words for each mask token and ranks the candidates by probability. This can be done because the model has been fine-tuned on domain-specific terminology, such as Reddit posts that talk about drugs. In the candidate word ranking, sentences that are domain-specific will have a lot of domain-specific words. But for generic sentences the candidates will be ordinary words that are unrelated to the domain. In order to check for specificity, only sentences where one of the target keywords appears in the top ranked candidate words are kept for analysis. Then, using the MLM again, the euphemism candidate list is generated by calculating the probability of a word candidate appearing in the sentence and returning all candidates ranked by weight.

This paper experimentally evaluates the functioning of their model and discusses potential for further use for it by running it on another domain. The euphemism identification stage of their model was not considered as their publicly released code did not include the results for identification. However, it is notable that Zhu et al state that their work on identification is the first of its kind.

Firstly, the drug-domain specific sample dataset along with the provided keyword and ground truth list were run for a comparison benchmark. The euphemism candidate list generated was the following:

> "**weed**, **coke**, **acid**, cannabis, alcohol, **speed**, mushrooms, pills, md, crack, **shatter**, **h**, **hash**, **crystal**, **pot**, **powder**, **x**, l, free, **bud**, product, **k**, oil, **hydro**, **wax**, pill, **bars**, **lucy**, psychedelic, rc, new, **lean**, narcotics, ir, **sugar**, **spice**, **stuff**, **blow**, something, gel, tobacco, fen, **e**, sample, synthetic, sh, **gold**, bad, met, cigarettes, products, shit, **pure**, 14, magnesium, heroine, crystals, **domestic**, legal, international, concentrate, substances, high, **dream**, mx, money, rocks, **poison**, **white**, 4, real, grade, bulk, quality, **ice**, **benz**, anything, **flower**, chemicals, **coffee**, bunk, sex, **blue**, **haze**, **green**, ether, pharmaceuticals, fe, tablets, good, vodka, supplements, guns, substance, mine, dom, cash, total, everything, lithium,"

The words in bold are words that are found in the euphemism ground truth list, and the underlined words are deemed incorrect euphemisms by themselves, but form part of a true euphemism. All other words are 'false positives', words that to the model appear to be euphemisms but are not contained in the ground truth list. By observing the false positives, it is possible to manually discover new euphemisms.

The top-10 precision is 0.50, which they assert is significantly higher than the next two best algorithms available. The 'p@k' evaluation method of an algorithm measures what percentage of the top k generated results are correct, in this instance 50% of the top 10 results being in the ground truth list.

Next, their model was run on a dataset of toxic Twitter discourse (Iyer, 2021). Iyer combined multiple available datasets that were imbalanced in their ratio of toxic to non-toxic tweets, in order to create a balanced-out dataset.

The aim of running it on this dataset was to detect when women were talked about in a hateful way. This is to potentially explore this kind of euphemism detection in the hate speech domain rather than illegal commerce. As hate groups find new ways to escape moderation and spread their ideology, keeping up by using methods specifically designed to detect covert insults could potentially be much more effective than other kinds of automated moderation such as toxicity scoring.

In order to run the model, a list of known offensive terms for women were collected and formatted to the requirements of their model. However, it appears that having only one target keyword as input was not sufficient for the model, as only 57 candidates were generated and all of them were filtered out as uninformative.

When the keyword list was modified to also include words for men (appendix 1), the following output was generated:

> "girl, guy, dude, **lady**, person, chick, one, shit, dick, pussy, wife, friend, vendor, fuck, boy, kid, student, cop, male, **female**, dog, ass, fool, asshole, bastard, teenager, teen, mother, child, **whore**, good, dealer, baby, thing, voice, cat, men, citizen, sub, lesbian, stranger, **bro**, ones, god, horse, people, cock, women, prince, boss, bastards, family, face, hat, partner, leader, mate, coat, couple, legend, character, boys, fucking, girlfriend, go, human, husband, bout, soul, joke, politician, bull, community, mom, guys, **prick**, brother, liar, government, father, resident, hog, rock, idiot, lot, user, bit, customer, gun, artist, racist, king, hero, nurse, house, snake, gentleman, troll, date, job,"

The achieved top-10 precision of 10% is significantly lower than the paper's top results. Moreover, 96.25% or 1925/2000 of euphemism candidates were discarded as uninformative, which is 3% higher than for the sample drug dataset. This is most likely due to the generic nature of the domain, and the small number of target keywords. It is also worth noting that 80 out of the 100 candidate words are nouns used to describe either men or women, so despite the domain being more generic, the model can propose candidates that fit the request.

In order to check these hypotheses, the same experiment was repeated with a larger hate speech dataset, which came with its own validated list of keywords. The dataset in question (Gomez, 2019) is a dataset of 150,000 extracted tweets that contain both images and textual information. The tweets were manually labelled by a team of volunteers with scores based on what kind of hate speech they contain, if any (Gomez et al., 2019). From their research findings, there is an observed overlap with hate speech and euphemistic speech: in order to express negative sentiment while evading moderation, new and subtle attacks are invented. As the aim of their research is not specifically to spot offensive terms but to compile tweets that attack communities, it potentially provides a sample where the confirmed set of keywords are used in a euphemistic sense.

For the purposes of this experiment, only the text data was needed. The textual information was extracted and cleaned using a custom script (appendix 2). Moreover, the keywords had to be modified to fit the formatting requirements of the input, including adding the answers (appendix 3). To get an answer list that would pass the model's input requirements, each keyword was linked to the nearest noun representative of the demographic group that the euphemism was referring to, and these had to be singular words. In this process some of the words had to be removed; words like "WomenAgainstFeminism" were clearly hashtags and did not fit the sentence structure requirements of the model, which checks for nouns.

The results were as follows:

"The percentage of uninformative masked sentences is 1863/2000 = 93.15%
[Euphemism Candidates]:
red, fucking, racist, american, old, little, blue, green, pink, random, homeless, yellow, big, dead, fat, street, good, college, new, bad, beautiful, plastic, poor, high, hispanic, urban, english, brown, dumb, indian, spanish, rich, one, chinese, german, city, drunk, mad, ass, blind, nice,

```
single, real, older, dark, russian, naked, local, dirty, religious, crazy, stupid, straight, sober,
ghetto, suburban, ugly, attractive, orange, free, purple, younger, grey, normal, human, lovely,
tight, african, best, right, girl, shit, smart, close, middle, japanese, full, fuck, looking, junk, cute,
internet, like, blonde, small, christian, cool, certain, female, hot, gold, pure, different, blacks,
tall, utter, european, greek, mail, wise,
Top-10 precision is (0.00, 0.00)
Top-20 precision is (0.00, 0.00)
Top-30 precision is (0.00, 0.03)
Top-40 precision is (0.00, 0.03)
Top-50 precision is (0.00, 0.02)
Top-60 precision is (0.00, 0.02)
Top-80 precision is (0.00, 0.01)
Top-100 precision is (0.00, 0.01)"
```

Although the dataset was confirmed to have contained the keywords, running the model on it gives precision scores of 0, or near 0. The only underlined word is “plastic”, because it is part of the word “plastic paddy” which is in the euphemism answer list for Irish. Empirically it also appears that the euphemism candidates are only loosely related to the domain of the target keywords. It can be observed, however, that a lot of origin-related words (“american”, “african”, “spanish” etc) appear on it, because the keyword list mostly contained attacks on origin.

It is possible that because the tweets were accompanied by images, some crucial context is contained in the images rather than in the text itself. Moreover, Gomez et al. add that due to the subjectivity of the task, it is possible that the accuracy of the annotations is directly related to how strong the hateful attack in the tweet is. In the context of this paper, the main concern thus is that although there is a perceived use of the insult keywords in a euphemistic way (a word is used to replace another word) in the dataset, it is difficult to confirm with their annotation method how many proper euphemisms the dataset may contain.

However, as demonstrated by the low accuracy of both the experiments, it appears that the main caveat is that without context-specific pretraining the model cannot produce reliable results. Zhu et al’s model has been trained in the context of vending and use of illicit products, whereas in hate speech the context generally is a targeted attack against a group of people or individual. These two domains are likely to have very different sentence structures. As pretraining the model was out of the scope of this paper, context-specific results could not be obtained. With pretraining and finetuning addressing the differences in context, it is likely that this model, or a similar approach, would be able to reach accuracy scores akin to the ones presented by Zhu et al.

Another caveat of this euphemism detection solution is that it does not consider multi-word euphemisms. In the domain of drugs or weapons, this may not be an issue, but in hate speech (as demonstrated by Appendix 3), many euphemisms are multi-word descriptors. This difficulty was demonstrated by the results only picking up one half of the euphemisms, like "plastic". It is likely that this resulted in some inaccuracy. By re-running it with the multi-word euphemisms removed, the results are marginally better with an average top-k precision of 0.2, but the euphemism candidates had very little variation from the previous results.

Zhu and Bhat (2021) address the problem of multi-word euphemism detection. The approach is akin to the single-word euphemism detection, with the difference that potential euphemism candidates are pre-selected using cosine similarity instead of an MLM, and instead of BERT (that can only rank single words), SpanBERT (Joshi et al., 2020, cited by Zhu and Bhat, 2021) is used to rank the euphemism candidates.

There is much room for improvement for an overarching approach to content moderation, where we may not know exactly what kind of unwanted euphemistic speech we are looking for. In the case of euphemistic speech targeted at groups of people, the experiments in this paper show that attempting to detect such instances may remain too vague for a model like this to accurately map out the euphemisms. Especially because this model discards such a large portion of the euphemism candidates as uninformative (even for the sample dataset), its uses as a standalone remain very domain specific. For the purposes of moderating illicit speech in general, there would have to be a suit of models all trained and finetuned on a specific style of euphemisms. While determining the scope of euphemisms and understanding the format of each is a promising first step, it seems unlikely that a platform would want to invest into a model that can only handle one kind of speech style. Illegal speech such as drug commerce is likely to take priority in moderation efforts.

As Zhu et al (2021) mention, their approach is well suited for a constant review-style approach as support for human moderation - their model does not as a standalone provide an answer for an automated or machine-based approach.

3.2 Ever-evolving communication

As seen in Gomez et al. (2019), most online communication involves formats other than text, such as images, videos, and emojis. Their research demonstrates that these different formats are not standalone, and it is not uncommon to have a combination of two or more formats interlinked in a single post. While they confirm that the text caption usually provides a good sense of what the overall intent is, which the Zhu et al. (2021) and Zhu and Bhat (2021) text-based approach would be suited to detect, the use of images and emojis as standalone communication are also used to convey euphemistic and hateful messages.

Hagen et al. (2019) state that due to the strict 280-character limit on Tweets, users rely heavily on emojis to convey standalone intent and to accompany text. In their research, they manually labelled tweets under the "#WhiteGenocide" hashtag as either favourable or unfavourable to white nationalism and collected data on what emojis were most frequently used by both groups. It is significant to note that in general, a single emoji does not carry universally recognised hateful meaning, and rather an analysis of both context and the individual ideology of the emoji user is needed to determine intent. This adds another layer of difficulty in hate speech detection.

Hagen et al. found supporters of white nationalist ideology to frequently use emojis of country names, highlighting both the international nature of the white nationalist movement as well as the nationalist sentiment it carries. Other emojis frequently used included a red X (❌), to indicate being "shadow-banned" (a user that is shadow-banned for violating code of conduct will not have their account or tweets recommended to other users), and the frog emoji (🐸). The frog emoji in particular carries significant meaning in the alt-right community, as it is a reference to the internet meme "Pepe the Frog", which has been reappropriated to symbolise their ideology.

Pepe the Frog's potential as a hate symbol comes from the fact that it is also a meme with more general appeal (Peters and Allan, 2021). This gives it contestability - an outsider might not always recognise that it is being used in a right-wing context. This makes it the perfect euphemism - members of the alt-right can signal to each other while remaining innocuous in the eyes of other social media users. The use of humour and absurdity are ideal for masking extremist ideology, as it blends in well with more general uses of the same meme, and thus have great deniability - even an openly racist iteration of Pepe can be dismissed as a meaningless joke that does not represent the poster's actual views. For in-group members, it "reaffirms boundaries of inclusion" and thus strengthens their stance against those that they are trying to exclude. The wide-scale appeal of Pepe is why to this day, despite being a known hate symbol, it is still being used by members of the alt-right. Because 'Pepe' symbolises such an elaborate set of views, it would be challenging to map out its meaning algorithmically, as it is not just a case of assigning a single word meaning to it. The same could be said for most memes and multimedia communications.

When deployed in content moderation, it is especially important that models perform well, as underperformance or inaccurate results have a direct impact on users. Kirk et al. (2021) created a suite of functional tests that test the emoji literacy of different models built for the purpose of hate speech analysis. They define seven distinct ways in hate speech where emojis replace an element of the sentence: threatening verb, attacked identity, descriptive noun, one character or part of a word, more than one word is swapped, a negative emoji accompanies a neutral statement, and a positive emoji accompanies a hateful statement. As seen, even on sentence - and word level the task of detecting emoji meaning is extremely case-specific and nuanced.

Their approach uses these definitions to manually generate test cases - fake statements that take the same form as real-life hate speech. They ran the test cases on DeBERTa (Ma et al., 2021, cited by Kirk et al., 2021) to determine how its performance differs between text-only hate speech and hate speech with emojis. Their results showed that although the number of misclassified emoji-text entries was over 60% initially, with training it lowered to about 40%. This is still higher than the 25% error rate DeBERTa has with text-only input after training.

This shows that although there is potential to train models to take emoji-input into account, there is still a gap in accuracy, and the inclusion of emojis is likely to make detecting intent more challenging. Their test suite is a promising start in considering the impact of non-textual communication on identifying gaps in content moderation models. They highlight, however, that their tests themselves only consider a constrained set of emojis and English language statements.

As a response to the constraints of language-specific models, Mubarak et al. (2022) detect offensive language and hate speech through emojis as a language-independent moderation method. This offers a way to capture illicit content without the need to consider the language or specific speech patterns. Although their research specifically focused on capturing Arabic tweets, they demonstrate that their emoji-based model also performs well in English. Some cultural differences in emoji usage, however, suggest that some finetuning may be needed depending on which language is being analysed.

They demonstrate that both Arabic-specific BERT models and multilingual BERT models can perform well when trained on their dataset, meaning that in future work there is potential to train a model that not only includes emojis in understanding illicit language, but may also use the emojis as anchors to understand offensive language in a way that transcends language barriers.

4. Interdisciplinarity

In this paper several ways in which natural language processing can provide solutions for content moderation and euphemism detection have been presented. However, as seen by the number of concepts needed to accurately define the problem of illicit online speech, computer science merely provides the tools, whereas multiple other disciplines are needed to determine what exactly needs to be addressed and how, making interdisciplinarity an inherent feature of this problem domain.

Understanding the motivations of the demographics that participate in euphemistic illicit speech on online platforms is crucial for defining the problem domain and collecting the data needed for training - in order to collect accurate data there needs to be a clear sense of what communities are in question (e.g. the alt-right) so that the associated forums, hashtags and keywords can be studied. This definition could be done using fields such as ethnography, sociology, or psychology.

The form and methods of euphemistic illicit speech could be defined through linguistic steganography. Chang and Clark (2014) define linguistic steganography as the practice of hiding information in natural language text. The aim of linguistic steganography is to transmit messages through an open channel that only the receiver will be able to understand or decode, in a form which appears innocuous to all other observers. One of the most common techniques of linguistic steganography is substituting synonyms. They propose a novel approach for synonym substitution for the purposes of encoding information, which could then later be decoded by a natural language processing algorithm. In their approach, every word is a vertex in a graph, encoded using a vertex coding algorithm, and connected to their synonyms by edges. By using this algorithmic approach, they create potential for a formalised approach to euphemism detection.

Developers of natural language processing systems also need clear guidance on what kind of models are needed, as raised by the Centre for Data Ethics and Innovation (2021). This guidance would directly come from the platform's internal policies, but in practice such decisions will often be influenced by government legislation.

Hence, a collaborative effort of disciplines is needed to accurately respond to this interdisciplinary problem.

5. A combined approach to automated online content moderation

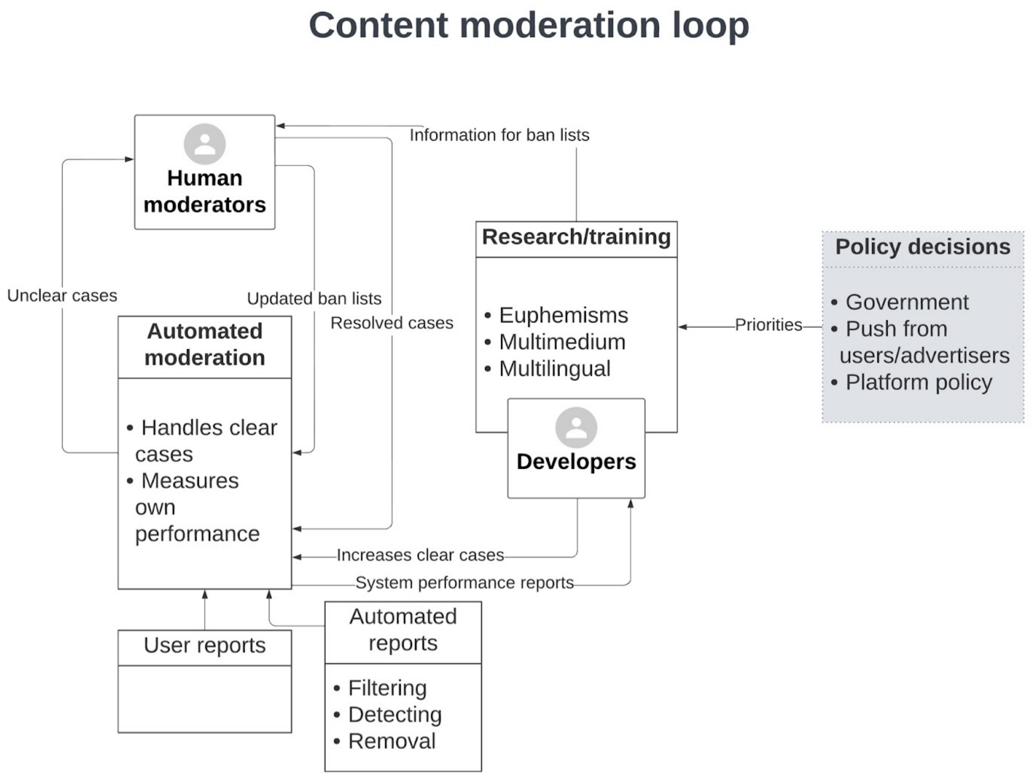

Figure 2: Proposal for a comprehensive content moderation loop

This proposal details the potential organisation of a content moderation loop that summarises all the aspects discussed in this dissertation. A multi-layered approach, where elements of the automation are separated, allows for a system that not only takes pressure off human moderators, but is also scalable and continuous. With the ever-evolving nature of online communications it is crucial that moderation indeed is a continuous loop, that constantly refreshes the model to be up to date with new ban-worthy terms and forms of euphemistic communication. The aim of this proposal is to demonstrate the full scope of the content moderation issue, as well as showcase that none of the solutions explored in this dissertation will work as a standalone and instead will be a valuable part in a larger framework.

At the heart of the loop is a high-quality automated moderation system, which would be able to independently resolve most moderation cases. By taking reports both from users and from a separate automated report system that combines the three methods of automated moderation, a constant review-style moderation would see most cases handled by the model with high accuracy. By measuring its own performance (such as the p@k precision metric in Zhu et al. (2021)), it would be able to classify which cases are straightforward, and which it is not able to classify accurately enough to make a decision.

The cases that are not precise enough would then be moved to human moderators, who will be able to make a decision that is shared with the automated system for future reference. Human moderators would also oversee maintaining and updating ban lists, which would be informed by continuous observation of new illicit and euphemistic words by a model akin to Zhu et al. (2021) and Zhu and Bhat (2021). Anomalies such as cases with unusually many user reports would also be mainly handled by human moderators.

In research, the main difficulties of speech moderation discussed in this dissertation would be covered. A model such as Zhu et al. (2021) and Zhu and Bhat (2021) that can handle both single-word and multi-word euphemisms with high accuracy would be beneficial in expanding standardised moderation. In addition, further research into detecting euphemistic illicit speech in multiple formats and languages would also be continuous. This proposal aims to explicitly emphasise that euphemisms are a separately complex problem from normal content moderation, as the intentional ambiguity adds multiple considerations that would have to be researched thoroughly to achieve accurate results. Rather than aiming to build a model that can do all of these things right away, deployment would be on an incremental basis, allowing research to keep up with the evolving nature of online communications.

Crucially, outside contributions such as government policy, pressure from advertisers and users, and the platforms' internal policy will determine what takes priority in content moderation. Research into what aspect of illicit speech is the biggest threat now will likely contribute into where moderation efforts will concentrate.

This dissertation recognises that such a content moderation proposal cannot with confidence be asserted as a fit-all solution, and it can only be asserted that in an ideal solution these elements would be in place. With social media platforms all being privately moderated and little to no information on their exact moderation methods being available, discerning how closely current moderation styles adhere to this model is not possible. It has been confirmed by the Centre for Data Ethics and Innovation (2021), however, that the major large platforms all prefer human moderation. Considering this, this proposal contests this by asserting that an ideal solution would involve automation where possible, as it has been demonstrated that natural language processing models can handle the task reliably.

It is also difficult to fathom that euphemisms would in the near future be formally regulated in the same manner as explicitly illicit speech. In the perspective of social media platforms urging for better guidance on what to regulate and their hesitance to overly infringe on free speech (Centre for Data Ethics and Innovation, 2021), euphemisms that do not directly pertain to illegal activity would likely not take priority in moderation efforts. It is also unlikely that platforms would change their moderation methods based on what a governing body dictates, as in general any legal frameworks simply lay out what kind of content must be removed and what the platforms' liability levels are.

This paper has demonstrated that natural language processing solutions such as the euphemism detection model (Zhu et al., 2021; Zhu and Bhat, 2021) are able to accurately respond to demands of modern content moderation, if the problem scope is clearly defined and appropriate sources for model training are available, and thus deserve a place in the larger content moderation framework.

Automated content moderation is not without its issues, however. The problem with training stems from the need for quality data. With a problem such as euphemistic speech, having large amounts of data to train from is necessary in order for models to determine patterns of speech. Such data will most likely have to be collected as part of the research, thus adding additional time and resource considerations. The larger issue in content moderation and euphemisms pertains to the definition of the problem, as well as the unclear nature of who has a say in content moderation, and how.

Future research would need to be conducted in how exactly a combined approach might be achieved algorithmically. The issues pertaining to specific language detection methods still need further research to evaluate their viability in real-life moderation. Additionally, research into how closely combined these methods would be in content moderation – whether they would all be under one algorithmic solution or rather a combination of multiple – would also be needed. Overall, the field of euphemism detection and its applications to content moderation have ongoing promising developments, and it can be hoped that in the future online content moderation sees more automated solutions being used.

6. Limitations

The scope of the experimental evaluation in this paper was limited by a lack of quality public datasets pertaining to the problem, a problem also raised in Zhu et al. (2021), who resorted to extensively collecting data from millions of online posts in order to meet their research needs. As the ethics limitations of this undergraduate dissertation did not permit for custom data scraping, a publicly available dataset was chosen, which entailed some limitations in regards to the domain being studied - as established in the methodology, there was no clear confirmation that the datasets chosen contained the kinds of sentence structures that enable euphemism detection as defined by Zhu et al.

(8409 words)

7. Appendix

7.1 Appendix 1

Custom keyword list for offensive terminology for women and men:

Woman: Babe; Becky; Bimbo; Bitch; Cougar; Cunt; E-girl; E-thot; Female; Femcel; Feminazi; Gold digger; Karen; Lady; Pick-me; Ran-through; Slut; Stacy; Thot; Whore

Man: Beta; Bro; Chad; Dickhead; Faggot; Himbo; Incel; Neckbeard; Prick; Simp; Soyboy; Whiteknight

7.2 Appendix 2

Custom script for tweet extraction:

```
import json, tempfile
import pandas as pd
import re
import fileinput
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize

with open('MMHS150K_GT.json', 'r') as mmhs:
    f = json.load(mmhs)

for i,j in f.items():
    text = j["tweet_text"] + "\n"
    text = text.lower()
    text = re.sub(r"(@[A-Za-z0-9_]+)|[^\w\s]|#|http\S+","",text)
    with open("tweet_text.txt", "a", encoding="utf-8") as d:
        d.write(text)
```

7.3 Appendix 3

Modified keyword list from Gomez (2019):

asian
arab
black
blasian
british
disabled
gay
jewish
mexican
muslim
white
woman
young

```
zionist
```

Modified euphemism answer list from Gomez (2019):

```
Asian: Chinaman
Arab: Camelfucker
Black: Coonass; Nigga; Nigger
Blasian: Bamboocoon
British: Limey
Disabled: Retard; Retarded
Gay: Dyke; Faggot
Jewish: Jsil
Mexican: Spic; Wetback
Muslim: Muzzie; Raghead
White: Hillbilly; Redneck; Rube; Whigger; Wigger; Wigerette
Woman: Bint; Bitch; Cunt; Feminazi; Sjw; Twat;
Young: Yobbo
Zionist: Zionazi
```

8. Bibliography

Bahdanau, D., Cho, K. and Bengio, Y. (2014). *Neural Machine Translation by Jointly Learning to Align and Translate*. [online] arXiv.org. Available at: https://arxiv.org/abs/1409.0473.

Bhat, P. and Klein, O. (2020). Covert Hate Speech: White Nationalists and Dog Whistle Communication on Twitter. *Twitter, the Public Sphere, and the Chaos of Online Deliberation*, pp.151–172. doi:10.1007/978-3-030-41421-4_7.

Burack, E. (2020). *A List of Antisemitic Dogwhistles Used By Donald Trump*. [online] Hey Alma. Available at: https://www.heyalma.com/a-list-of-antisemitic-dogwhistles-used-by-donald-trump/ [Accessed 6 May 2022].

Centre for Data Ethics and Innovation (2021). *The role of AI in addressing misinformation on social media platforms*. [online] Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1008700/Misinformation_forum_write_up__August_2021__-_web_accessible.pdf.

Chang, C.-Y. and Clark, S. (2014). Practical Linguistic Steganography using Contextual Synonym Substitution and a Novel Vertex Coding Method. *Computational Linguistics*, 40(2), pp.403–448. doi:10.1162/coli_a_00176.

Criddle, C. (2021). Facebook moderator: 'Every day was a nightmare'. *BBC News*. [online] 12 May. Available at: https://www.bbc.com/news/technology-57088382.

Danskin, I. (2019). *The Alt-Right Playbook: How to Radicalize a Normie*. *YouTube*. Available at: https://www.youtube.com/watch?v=P55t6eryY3g.

Gillespie, T. (2018). *Custodians of the internet : platforms, content moderation, and the hidden decisions that shape social media*. New Haven: Yale University Press.

Gomez, R. (2019). *Multimodal Hate Speech*. [online] www.kaggle.com. Available at: https://www.kaggle.com/datasets/victorcallejasf/multimodal-hate-speech [Accessed 8 Apr. 2022].

Gomez, R., Gibert, J., Gomez, L. and Karatzas, D. (2019). *Exploring Hate Speech Detection in Multimodal Publications*. [online] Available at: https://arxiv.org/pdf/1910.03814.pdf [Accessed 19 Aug. 2021].

Hagen, L., Falling, M., Lisnichenko, O., Elmadany, A.A., Mehta, P., Abdul-Mageed, M., Costakis, J. and Keller, T.E. (2019). Emoji Use in Twitter White Nationalism Communication. *Conference Companion Publication of the 2019 on Computer Supported Cooperative Work and Social Computing*. doi:10.1145/3311957.3359495.

Hindman, M. and Barash, V. (2018). *Disinformation, 'Fake News' and Influence Campaign*. [online] Knight Foundation. Available at: https://s3.amazonaws.com/kf-site-legacy-media/feature_assets/www/misinfo/kf-disinformation-report.0cdbb232.pdf.

Iyer, A.U. (2021). *Toxic Tweets Dataset*. [online] kaggle.com. Available at: https://www.kaggle.com/ashwiniyer176/toxic-tweets-dataset.

Kirk, H.R., Vidgen, B., Röttger, P., Thrush, T. and Hale, S.A. (2021). Hatemoji: A Test Suite and Adversarially-Generated Dataset for Benchmarking and Detecting Emoji-based Hate. *ArXiv*. [online] Available at: https://arxiv.org/pdf/2108.05921.pdf [Accessed 7 May 2022].

Land, M.K. (2019). *Against Privatized Censorship: Proposals for Responsible Delegation*. [online] papers.ssrn.com. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3442184 [Accessed 5 May 2022].

McCulloch, L. and Furlong, S. (2019). *DM for Details: Selling Drugs in the Age of Social Media*. [online] *Volteface*. Available at: https://volteface.me/publications/dm-details-selling-drugs-age-social-media/.

Mozilla Foundation (2021). *YouTube Regrets*. [online] *Mozilla Foundation*. Available at: https://foundation.mozilla.org/en/youtube/findings/.

Mubarak, H., Hassan, S. and Chowdhury, S.A. (2022). Emojis as Anchors to Detect Arabic Offensive Language and Hate Speech. *arXiv:2201.06723 [cs]*. [online] Available at: https://arxiv.org/abs/2201.06723 [Accessed 7 May 2022].

Peters, C. and Allan, S., 2021. Weaponizing Memes: The Journalistic Mediation of Visual Politicization. *Digital Journalism*, 10(2), pp.217-229. Available at: https://www.tandfonline.com/doi/full/10.1080/21670811.2021.1903958

Putnam, R.D. (2000). *Bowling alone: the Collapse and Revival of American Community.* New York: Simon & Schuster.

Saeed, M. (2021). *An Introduction To Recurrent Neural Networks And The Math That Powers Them*. [online] Machine Learning Mastery. Available at: https://machinelearningmastery.com/an-introduction-to-recurrent-neural-networks-and-the-math-that-powers-them/.

Saul, J. (2018). Dogwhistles, Political Manipulation, and Philosophy of Language. *New Work on Speech Acts*. [online] Available at: https://oxford.universitypressscholarship.com/view/10.1093/oso/9780198738831.001.0001/oso-9780198738831-chapter-13.

Sayce, D. (2020). *Number of tweets per day? | David Sayce*. [online] David Sayce. Available at: https://www.dsayce.com/social-media/tweets-day/.

Speckhard, A., Ellenberg, M., Morton, J. and Ash, A. (2021). Involuntary Celibates' Experiences of and Grievance over Sexual Exclusion and the Potential Threat of Violence Among Those Active in an Online Incel Forum. *Journal of Strategic Security*, 14(2), pp.89–121. doi:10.5038/1944-0472.14.2.1910.

Twitter Help Center (2019). *The Twitter Rules*. [online] Twitter.com. Available at: https://help.twitter.com/en/rules-and-policies/twitter-rules.

United Nations Office on Drugs and Crime and United Nations Counter-Terrorism Implementation Task Force (2012). *The use of the Internet for terrorist purposes*.

Zhu, W. and Bhat, S. (2021). *Euphemistic Phrase Detection by Masked Language Model*. [online] Available at: https://arxiv.org/pdf/2109.04666.pdf [Accessed 6 May 2022].

Zhu, W., Gong, H., Bansal, R., Weinberg, Z., Christin, N., Fanti, G. and Bhat, S. (2021). Self-Supervised Euphemism Detection and Identification for Content Moderation. *arXiv:2103.16808 [cs]*. [online] Available at: https://arxiv.org/abs/2103.16808 [Accessed 8 Apr. 2022].