DEPARTMENT OF PHILOSOPHY,
LINGUISTICS AND THEORY OF SCIENCE

IDENTIFYING HATE SPEECH IN SOCIAL MEDIA THROUGH CONTENT AND SOCIAL CONNECTIONS ANALYSIS

Milan Stanišić

Master's Thesis:	15 credits
Program:	Master's Program in Language Technology
Level:	Advanced level
Semester and year:	Spring, 2023
Supervisor:	Aleksandrs Berdicevskis
Examiner:	Simon Dobnik
Keywords:	hate speech, social media, natural language processing, classification

Abstract

Hate speech is a problem which puts its targets at risk of serious harm. It spreads fast and has a real influence on the society because of the ubiquity of the internet and social media, and so various research efforts have been put to find solutions to automatic hate speech detection. Despite major developments in the field, challenges with data scarcity and characteristics often cause solutions reported in previous research to overfit the datasets that were used to train and test them, which results in dramatic performance losses and failures in generalization. This study addressed this issue, it tried to find a solution that would mitigate overfitting effects originating from these issues and enhance language-based classifier with extra user information concerning one's social connections. It compared two single-source models – one based on textual information, and the other based on information concerning one's social connections and proposed a joint decision engine that selects the model whose class assignment was more certain for a given instance. Although the single-source models' performance dropped drastically on test data, the joint decision engine succeeded in reducing some of the issues related to overfitting, improving the overall performance. This observation suggests that simple solutions might be efficient in reducing model overfit and paves the way towards validating these findings.

Contents

1 Introduction

Communication of today's world is largely dependent on the internet which is nowadays becoming ubiquitous and accessible for almost everyone in the western civilization. This information exchange tool allows nearly all of its users, regardless of their origin and beliefs, to contribute to miscellaneous discussions and be heard by others with almost no effort. The pace of information spread is staggering as through connections between people, one message can reach thousands if not millions of users within a blink of an eye (Kovács et al., 2021). This encourages internet users to freely express themselves on virtually any topic and publicly convey their attitude towards various subjects, either positive or negative (Kovács et al., 2021; MacAvaney et al., 2019). Additionally, the internet provides a certain degree of anonymity which is desirable by its users not only to hinder linking an online account to who they really are but also to create fake personalities so that various online groups with restricted access can be reached by the user (Kang et al., 2013). All these factors: ubiquity, freedom of speech and anonymity make the internet an attractive place to get involved into a wide range of topics with a low risk of potential consequences.

However, there is a major flipside of these features of online platforms. Since the internet allows users to be anonymous, they feel like no one can draw any consequences from what they post, which leads to an abuse of the freedom of speech (MacAvaney et al., 2019). This in turn leads to a spread of hate speech which, although its definition is vague and varies across researchers and parties, can be roughly described as a speech act that is intentionally harmful towards a certain societal group (Sellars, 2016). This offensive content is particularly used by extremists and appears frequently across discussions touching on things one cannot control such as race or religion and controversial topics like abortion or political preferences; this is because extremists tend to strongly defend their claims, oftentimes using abusive and hateful language against their opponents (Ribeiro et al., 2020). Online platforms, such as YouTube, can become a place where extremism against given social groups or issues is rising ad spreading (Ribeiro et al., 2020).

Extremists' claims can persuade people whose opinions on touchy topics are not as radical, leading to the spread of extreme opinions (Ramos et al., 2015; Ribeiro et al., 2020). According to Ribeiro et al. (2020), this phenomenon of extremism spread is particularly dangerous as the internet is becoming a primary information source, leading to possible misinformation and hatred spread outside the online world. In the most extreme scenario, over time, it can eventually bring social movements and ideologies based on exclusion of certain groups such as white supremacy back to life (Ribeiro et al., 2020). Moreover, it can influence the outcome of various legal and democratic processes (Arango et al., 2019). Therefore, since use of hate speech can characterize extremists – defenders of values whose magnitude on the opinion spectrum markedly outlies from the population's general view who are prone to promote violence against their opponents (Govers et al., 2023) – in order to alleviate the phenomenon of extremism spread, it is necessary to identify haters, i.e. those that actively use offensive language against others, on the internet.

Identifying hate speech – one of key indicators of dangerous extreme movements – and its authors is not a straightforward task. Manual labeling of hate speech is out of question due to the abundance of data on the internet uploaded daily (Arango et al., 2019; Badjatiya et al., 2017) – there is no way human annotators can keep up with the sheer volume of content that needs to be verified for the presence of hate, and so automatization is needed. However, automatization also faces challenges, mostly linked to the lack of consensus as to what hate speech actually is (Arango et al., 2019; MacAvaney et al., 2019). Hate speech is often a subjective matter; while some content might be perceived as hateful by one individual, some other individual might not consider the same content as such (Sellars, 2016). This does not allow defining its consistent definition across different parties and researchers, which causes problems with data annotation, an absolutely necessary step in introducing automatization (Founta et al., 2018; Kovács et al., 2021).

To solve these problems, researches use language models trained on datasets that are a result of merging datasets from previous researches or that were annotated by crowdsourcing, i.e. multiple people deciding

on the eventual label to reduce the subjectivity (Founta et al., 2018). However, there is yet another issue, probably the most relevant one when it comes to language-based hate speech detection. Entries on social media are usually short, as many platforms (e.g. Twitter) where hate speech is noticeably spread have put constraints on the posts' length (Tsourougianni & Ampazis, 2013). Additionally, users tend to use abbreviations that are not understood by language models to overcome this limitation. This in turn, as Tsourougianni & Ampazis (2013) emphasize, can cause language-based classifiers to produce unreliable results since data available in each instance are limited. Considering that some users post relatively short texts, even if the language-based hate speech classification model is well-built, this problem imposes the necessity to use additional user data to determine whether or not someone is a hater.

One way to classify someone as a hater is analyzing their connections with other users, an approach investigated by Mishra et al. (2018). As Mishra et al. (2018) explain, it is possible to profile a user and assess whether or not they produce abusive content by analyzing their social graph on the social network and check who they follow. This is because of the phenomenon of homophily, which states that people tend to form clusters with those who share similar beliefs and lead similar lifestyles, one is more likely to be a part of a certain group if they behave like its members (McPherson et al., 2001; Mishra et al., 2018). The impact of this phenomenon is visible not only among real life interactions like children and their decisions regarding who to play with (McPherson et al., 2001) but also on social platforms like Twitter (Mishra et al., 2018). This implies that haters do form clusters and groups since their behavior is similar, and that information on cluster membership might constitute an additional point of reference aside from the output of a language-based classifier. However, building a solid social graph often requires additional information on the users' engagement in distinct groups as one might not be involved in all groups uniformly (Del Tredici et al., 2019). These data, as MacAvaney et al. (2019) highlight, is not always accessible and its usage might be impossible for external researchers working on new solutions.

The question is whether it is possible to develop a system that could exploit data on a certain user available to all other users, such as post contents and followers lists, to alleviate issues that contemporary language models face when it comes to overfitting which has been proven to be problematic with the state-of-art solutions (Arango et al., 2019). Inspired by the work of Del Tredici et al. (2019) and findings of Arango et al. (2019), the research described in this thesis addresses the problem and verifies if the decision of a language model as to whether or not someone is a hater can be corrected using the decision of a model analyzing one's connections with other users, especially when the prediction of the language-based classifier makes predictions with a greater degree of uncertainty. It also verifies which classifier: language-based or social connection-based performs better. Therefore, this research instantiates the following research questions:

RQ_1: Are there any differences in performance between the single-source models?

RQ_2: Is it possible to combine those two models into one decision engine to improve the setup performance?

RQ_1 imposes the following hypotheses to check:

- H_0: there are no significant differences in the performance of both models.
- H_A: there are significant differences in the performance of the models checked.

In the case of RQ_2, the hypotheses are as follows:

- H_0: the decision engine proposed does not produce better results than the single-source models on their own.
- H_A: the decision engine improves the metrics of the setup.

The next section describes in-depth previous research touching on related matters and provides additional background information that further motivates this research. Further sections describe in-detail the methodological approach, justify the choices made, and interpret the results of this experiment. Although the results of this experiment suggest that the decision engine is able to catch a smaller fraction of hate speech than the language model, such a solution improves the general performance of the setup in terms of the degree of randomness and reduces the false positive (FP) rate. This might be beneficial in terms of creating a setup that reduces the effects of overfitting which was commonly encountered in previous researches (Arango et al., 2019) and which was unwillingly observed in this project as well.

2 Related Work

2.1 Hate Speech: Definition and Spreading Mechanism

As aforementioned, there is no consensus when it comes to defining what hate speech actually is. This mostly has to do with the subjectivity of the matter, which makes the definition of this term vary from across researchers and organizations (Sellars, 2016). MacAvaney et al. (2019) gathered multiple definitions of hate speech to establish their common points and highlight differences between them. The key characteristic all of them were consistent about was that hate speech is an attack on the basis of some traits possessed by the target. Sellars (2016) further specified that the feature distinguishing a certain group or individual is usually something whose beholder cannot control, such as ethnicity, religion or sexual orientation, which makes any act of such an attack painful and diminishing for the target recipient. Some differences across various attempts to explain what this phenomenon is include the addressee – while some variants say hate speech is targeted towards an individual, others say it is aimed against a whole group of people (MacAvaney et al., 2019; Sellars, 2016).

These acts of verbal abuse aimed at certain people, groups or organizations are generally characterized by an extremely abusive and violent nature of their language (Sellars, 2016). One can thus infer that hate speech must thus be extremely negative in terms of its sentiment. However, this does not necessarily have to be the case and such an approach to defining hate speech is not accurate. This is because hate speech can in certain cases take form of an appraisal and support of some group or organization that promotes violent behaviors against some group/individual of interest (MacAvaney et al., 2019). Such cases would therefore be mistakenly classified as non-hateful. Additionally, Matalon et al. (2021) observed that some messages posted and shared in social media can get their sentiment inverted when put in a different context. This can potentially lead to indirect hate speech that interprets a positive message in a negative way; as a result, the original message can be accidentally misclassified as hate speech. Therefore, although hate speech is in most cases associated with abusive language that promotes violence (Sellars, 2016), there are situations where its sentiment might contradict this claim, and so identifying hate speech sheerly via sentiment analysis is insufficient.

Gathering all the information, hate speech can be said to be an attack whose goal is to hurt the recipient in as many aspects as possible, sometimes even prompting physical acts against the target. The target usually is chosen on the basis of some characteristics they were born with and that cannot be changed such as ethnicity, religion or sexual orientation, which makes hate speech closely related to extreme alt-right movements – an ensemble of strong and controversial opinions that claim supremacy of people who belong to a certain race and adhere to values considered traditional in a given society (Ribeiro et al., 2020; Sellars, 2016). As Sellars (2016) notices, hate speech acts often promote violence, which further highlights their dangerous nature.

By saying that violence is the only way of dealing with the targeted group/individual, haters can lead to actually committing violent acts in the real world by themselves or their adherents and marginalizing targets in the society (Ribeiro et al., 2020; Sellars, 2016). It might even be the case that, as Arango et al. (2019) mention, hate speech can spread to such a degree that it influences the external world of politics, which can harmfully impact targeted groups legally and politically. These characteristics of hate speech make it related to two closely related terms – extremism and radicalization – that are centered around similar concepts (Govers et al., 2023; Ribeiro et al., 2020). While extremism stands for holding opinion whose strength is outstandingly large compared with the population, radicalization is a process during which one's viewpoint becomes more and more extreme (Govers et al., 2023). Given the relatedness of hate speech to these two key notions, it becomes apparent that it might have detrimental effects on the society.

Hate speech on online platforms spreads fast, and there are several phenomena underlying its staggering pace, mostly concerned with the characteristics of social networks. Since the internet allows joining virtually

any social circle, one's social network can grow substantially, as people become connected to more and more individuals which allows any opinion to be heard by a greater social circle (Ribeiro et al., 2020). This convenience of information sharing internet platforms offer, connected with the small world phenomenon described by Watts (1999), allows even faster information spread than it would be the case without internet as more each person, thanks to online platforms, is connected to more clusters. Even without the internet, the information would spread fast as according to Watts (1999), any person is connected to every member of the whole society through just a few intermediate connections, and the ease of opinion sharing offered by the World Wide Web seems to catalyze the spread due to a larger number of connections across individuals (Ribeiro et al., 2020).

The spread of hate speech, just like extremism and radicalization, can also be caused by the recommendation algorithms used by social platforms to make the user access content tailored to their points of interest. Ribeiro et al. (2020) explained this phenomenon through analyzing how YouTube users switch to racist and hateful alt-right content over time. They concluded that one of the key reasons why users of this platform are becoming more and more extreme is in the dynamics of users' activity. People holding extreme views tend to be persuasive and they strongly defend their claims, which can be visible through one's vivid activity caused by the need of substantiating their opinions and beliefs (Ramos et al., 2015; Ribeiro et al., 2020). As Ribeiro et al. (2020) noticed, this causes the extreme and dangerous content to exhibit an extremely high activity of users watching it, which is reflected in the ratio of comments per view. This in turn can make the algorithms decide that it is a good-quality content because it attracts a lot of attention, which results in recommending it to users whose preferences are also conservative but not as hateful and extreme as those presented in that content (Ribeiro et al., 2020). And since among those extreme communities there is a feeling of reassurance that the claims supported by these groups are the only right ones, users gradually get fed with more and more extreme content, and are more and more strongly convinced that they are right (Ribeiro et al., 2020).

Ramos et al. (2015) pointed out yet another factor that contributes to the spread of extremism which, as Govers et al. (2023) mention, can be to a certain degree distinguished by characteristics comparable to those of hate speech, many of which are associated with expressing hostility towards the target group whose characteristics are opposed to the hater's group. After analyzing the ratio of extreme opinions on various matters among numerous communities across several years, they found out that the increase in the share of extreme opinion beholders is linked to some triggering events like an economic crisis which makes more people dissatisfied with the status quo of their lives. However, this is not the only necessary factor as it is also required that the population contain at least some individuals who already hold extreme opinion on the matter of interest and whose claims become convincing at the time the touchy issue is raised by the key event (Govers et al., 2023; Ramos et al., 2015). Therefore, one can infer that hate speech becomes more apparent and spreads more easily in similar circumstances.

2.2 Goals Behind Labeling Hate Speech

All of the characteristics of hate speech described above along with the ease with which it spreads present how it threatens the safety of the society, especially when it comes to those members of it who are more prone to be the targets of hate speech because of their personal characteristics. They highlight why identifying hate speech is necessary and show potential consequences of not doing so, and they allow listing some key goals of such a classification. The primary goal behind identifying and labeling hate speech, as well as its authors, is to reduce the pace with which it spreads (MacAvaney et al., 2019; Wich et al., 2021). Since this phenomenon can spread through recommendations on social media, this goal can be achieved via refraining identified hate speech from appearing in user recommendations (Badjatiya et al., 2017; Ribeiro et al., 2020). This way, users would not be recommended sources and authors that are labeled as hateful, which should slow down the pace with which hate speech spreads on the internet as access to such content would be hindered.

Humans, however, are not the only ones that are put at risk of being influenced or targeted by hate speech. As Badjatiya et al. (2017) mention, hate speech can also affect contemporary language models that are trained on content scraped from various social media sites. This is particularly concerning when taking various kinds of conversational agents into account where training data quality matters. An example of a considerable influence of hate speech is described by van Rijmenam & Schweitzer (2018) who brought a chatbot developed by Microsoft corporation for Twitter that was supposed to tailor its behavior to the users it interacted with. It was taken down just a few hours later because as it was exposed to hate speech from the side of its users, it started producing hateful and racist content. Through this example, van Rijmenam & Schweitzer (2018) highlighted the importance of training algorithms on unbiased data. Filtering hate speech before passing language content as training data to conversational agents and other language-based models has a potential of making these algorithms less biased, and this is another reason why the problem of hate speech identification needs to be tackled efficiently (Badjatiya et al., 2017).

Additionally, it is pivotal that this hate speech detection process be automated. The main motivation behind this is the immense volume of data that are uploaded to the internet daily; skimming through these resources in the search of hate speech is physically unfeasible using human labor force (Arango et al., 2019; Kovács et al., 2021). This need is also strengthened by the aforementioned fact that due to no consensus across researchers when it comes to defining hate speech, manual annotation is prone to subjectivity (Founta et al., 2018). Such a subjectivity can result in failing to capture some hate speech instances that one did not consider hurtful due to personal point of view. Usually, multiple annotators are employed to avoid this human bias through voting on what is the most appropriate label but this requires more human force, not to mention time resources (Founta et al., 2018; Kovács et al., 2021). Automatized techniques overcome these challenges as not only are they able to process the immensity of data uploaded daily within reasonable time but also they are less likely to be subjective as they their main goal is to find characteristics of data that are generalizable to all instances (Goodfellow et al., 2016; Kovács et al., 2021). Although automatization still requires some manual work with training dataset preparation, it eventually makes the task of hate speech classification scalable and doable.

2.3 Contemporary NLP-Based Approaches to Hate Speech Classification

One of the simplest approaches to automatizing hate speech detection relies on keywords related to hateful content such as curse words related to one's personal characteristics (Pereira-Kohatsu et al., 2019; MacAvaney et al., 2019). Although solutions based on this kind of word matching provide an easy to understand decision engine, they have several drawbacks that make them unpreferred in contemporary hate detection systems. First of all, keyword-based algorithms fail to catch those instances of hate speech where none of the keywords are used (MacAvaney et al., 2019). Additionally, since the selection of keywords is performed by humans, this approach brings drawbacks related to keyword choice as selecting only some certain words expressing hatred can restrict the whole system to be sensitive to only a fraction of all groups affected by hate speech (MacAvaney et al., 2019). For example, selecting words related solely to one's ethnicity will result in algorithm's inability to detect hate speech acts aimed at LGBT people. Manual selection of keywords also disregards the fact that words can change meaning depending on the context they are used in which can lead to many false positives generated by the system (Pereira-Kohatsu et al., 2019; MacAvaney et al., 2019). Therefore, more up-to-date solutions are not entirely based on keyword matching; instead, more sophisticated techniques are used.

A vast part of currently used solutions to automatizing hate speech detection seems to be based on various techniques that put Machine Learning (ML) tools in practice. Badjatiya et al. (2017) claim these techniques overcome human bias in determining what is a relevant feature of hate speech and what is not. They can capture different forms hate speech can take without the need of human intervention in the process – all that suffices is data and the classifier determines what causes a given instance to be considered hateful automatically. Such solutions resolve the issues mentioned by MacAvaney et al. (2019) with limitations

of using keywords and manual annotation. Pereira-Kohatsu et al. (2019) provided a thorough overview of studies on identifying hate speech using ML techniques that focus primarily on language characteristics, showing that hardly are other approaches to this problem used nowadays. They also showed the diversity of algorithms that can be applied in handling the task of detecting hate on the internet.

Some of the most popular ML algorithms used in hate speech detection include Support Vector Machines (SVMs) that are capable of distinguishing data with non-linear boundaries between two classes (Goodfellow et al., 2016; Pereira-Kohatsu et al., 2019). Other widely used techniques include neural networks and decision trees-based approaches, and some of them constitute a hybrid of these two approaches (Badjatiya et al., 2017). With their flexibility, neural networks overcome the problems faced by many classic ML methods with data non-linearity that is particularly common in language (Badjatiya et al., 2017). In addition to that, through backpropagation, they can improve themselves, which makes them better than classic Machine Learning algorithms (Goodfellow et al., 2016). In natural language data preprocessing, neural networks can also be used to extract abstract linguistic features that can than be encoded in embeddings and used to train task-specific classifiers (Jurafsky & Martin, 2021).

Speaking of embeddings – numerical vectors representing a given concept or a document that contain semantical information (Jurafsky & Martin, 2021) – there are also multiple ways they can be generated. Any textual document can be represented using such a vector, and the characteristics encoded within this kind of numerical representation can later be used to categorize text and speech, treating vector values as features or dimensions (Jurafsky & Martin, 2021). Basic methods involve computation of TF-IDF values or even simple frequency rates for words that appear in the data and putting them together into a vector (Badjatiya et al., 2017; Jurafsky & Martin, 2021). While the former technique usually outperforms the latter as it can assess the importance of a given term in the context it is found in, their major drawback is that vectors they produce can be large and sparse, which is computationally inefficient (Jurafsky & Martin, 2021). Therefore, more sophisticated techniques of embedding generation are used, and they are proven to cause hate speech classifiers perform better (Badjatiya et al., 2017). These include such techniques as FastText proposed by Bojanowski et al. (2017), which is described in-detail in Section 3. Combined with contemporary classifiers, such embeddings can help achieve satisfactory results (Badjatiya et al., 2017).

Previous studies have reported achieving good and robust performance. According to the overview provided by Pereira-Kohatsu et al. (2019), in some hate speech-related tasks, models like SVMs can reportedly attain impressive performance metrics on test sets, with precision and recall scoring up to 0.97. Another example is the model developed by Badjatiya et al. (2017) whose precision and recall both equal to 0.93; their proposed architecture was a decision tree-based model that used gradient boosting, a technique closely related to backpropagation that reduces the model's error over time through operations that aim to reduce the value of the loss function (Goodfellow et al., 2016; Ye et al., 2009). This model, along with randomly trained word embeddings, outperformed other algorithms on the dataset used and this achievement was highlighted in multiple other researches (Pereira-Kohatsu et al., 2019). However, it turns out that the models proposed in previous studies might not be as generalizable and robust as they seem to be on the basis of metrics reported, especially when new data are introduced, which causes performance metrics such as precision and recall to go down by even thirty percent (Arango et al., 2019).

2.4 Challenges of Language-Based Approaches

Despite impressive developments in the field, all contemporary solutions suffer from similar weaknesses, many of which are related to data quality, quantity, diversity and availability. Arango et al. (2019) highlight the issue that was underestimated by many researchers, namely the problems that arise when the classifier is trained on a dataset with various limitations. This is generally concerned with poor generalizability of the models and difficulties with applying them to new data as their training is usually restricted to some specific structure of the data imposed by its source (Arango et al., 2019). For example, user entries on Twitter

have a limitation on their length set, which restrains the models from generalizing onto a broader linguistic context (Tsourougianni & Ampazis, 2013). To provide a further justification for this claim, Arango et al. (2019) trained multiple models trained by previous researches, including Badjatiya et al. (2017), on datasets that differed from those used in the original studies to evaluate those classifiers. It turned out that none of the models tested, even the apparently sound model built by Badjatiya et al. (2017), made good predictions on unseen data instances as their metrics dropped substantially compared with the results presented in the original papers. This suggests that the problem of overfitting is common across models that primarily rely on analyzing textual content to make predictions on hate speech.

Data scarcity and problems with their quality are partially the reason why models overfit, i.e. perform poorly in classifying unseen instances. Founta et al. (2018) touched on the latter problem and state that apart from the lack of agreement across researchers when it comes to defining hate speech is a problem with consistency when it comes to labeling data. Not only is hate speech definition a subjective matter but also some labels are used interchangeably: for example, racist is often confused with hateful (Founta et al., 2018). Additionally, given that multiple researchers (e.g. Founta et al. (2018), Pereira-Kohatsu et al. (2019), Kovács et al. (2021)) mentioned that many algorithms, including those trained by themselves, are built on commonly used datasets that contain just a few thousand entries, the problem of generalization onto all possible hate speech scenarios becomes apparent. To make matters worse, it can be the case that multiple entries in the dataset originate from merely a few users, which further downgrades the ability of the models to generalize on new data because language content coming from one user is consistent in style and target (Arango et al., 2019). Combined with the fact that datasets based on social media channels like Twitter where data quality is questionable due to considerable length constraints (Tsourougianni & Ampazis, 2013), all these issues contribute to model overfitting and unsatisfactory results.

Arango et al. (2019) specified that even if data are enriched with additional instances, particularly those labeled as hateful due to their general scarcity, the issues with overfitting are alleviated but not resolved. The problem of data imbalance is not helpful regarding this matter – a small ratio of hate speech to non-hate speech instances causes the models to not generalize well on data as there are not enough positives (i.e. hate speech labels) for models to learn characteristics of hatred in text (Kovács et al., 2021). All of this suggests that pure text data might be insufficient in correctly classifying hate speech, and that more than just linguistic content is needed, particularly when data quality is questionable due to their length and amount (Arango et al., 2019; Del Tredici et al., 2019; Mishra et al., 2018). This imposes the necessity to explore additional information on users to determine whether or not they are haters with a greater degree of certainty.

2.5 Solution: Extra User Data

Mishra et al. (2018) saw room for improvement in classifying abusive language through applying principles of homophily. According to McPherson et al. (2001), homophily is a term referring to a phenomenon of people forming groups primarily with people who cherish similar values and have mutual points of interests. The bases of distinctions between groups are broad, ranging from the origin and socioeconomic background to political preferences (McPherson et al., 2001). Such groups are therefore uniform, which makes people not matching their characteristics rarely constitute their part (McPherson et al., 2001). As evidence shows, people form clusters according to this principle not only in the physical world, but also on social media platforms, which was observed by Zook (2012) who noticed that haters commenting on the U.S. presidential elections of 2012 were connected to each other rather than distributed evenly across the society. Additionally, as Dobnik et al. (2022) mentioned, people interacting with each other seem to use similar language style through semantic alignment, and so it might be that people using similar forms of hate speech are more likely to interact. Mishra et al. (2018) used this knowledge to build a model that apart from analyzing the one's language, analyzed also one's neighborhood in the network to compute an embedding representing the given user. Their results showed that information on social relations with other

users can indeed help in identifying authors of Tweets whose nature is abusive.

Using information on social connections also provides additional insights on the situation context. According to Dobnik et al. (2022), it is often the case that even whole phrases might change their meaning according to the context of a given situation. Since social connections of a given person more or less reflect the world that person is surrounded with (Mishra et al., 2018), information on them can be used to define the most probable context of a given utterance. They can be compared to the observations of the real world mentioned by Dobnik et al. (2022) as they help in contextualization. This is because social connections can explicitly present which social circles one is active in, reducing the number of possible contexts, they can be compared to selecting domains associated with a given person.

The idea of using social connections to better classify users into several groups is not new. Tsourougianni & Ampazis (2013) analyzed social connections between users to tailor follow recommendations to the structure of one's social graph. Although the goal of the task tackled by Tsourougianni & Ampazis (2013) is different from detecting haters and hateful content they produce, the principle stays the same: users with similar attitudes stay together. This is reflected in who follows whom as connections on social media do not necessarily represent actual friendships; apart from that, they can also show what the given user is interested in (Tsourougianni & Ampazis, 2013). Given this fact and observations made by Zook (2012) on the way haters form clusters on the internet, this implies that it is possible to identify such users using information on their connections. This was confirmed by Wich et al. (2021) who conducted an experiment using information on social connections aside from linguistic content produced by users. They analyzed if the fact that a user follows some accounts annotated as hateful might be useful information in hate speech detection task, and they managed to build a model that achieved satisfactory performance on a custom dataset.

However, hardly ever are people members of only one social circle, they usually contribute to multiple groups at once to various degree. According to Del Tredici et al. (2019), this poses a challenge for social connection-based approaches as it is never the case that people treat all groups evenly; rather than that, they contribute to some of them to a greater extent than to others. This fact is often omitted in approaches focused on social connections, and so Del Tredici et al. (2019) focused on taking that factor into account. In their experiment, they computed embeddings that represented social relations of a given person on the basis of the importance of connections between the users. The importance of such connections can be determined by analyzing one's behavior through retweeting and following other users on the platform (Del Tredici et al., 2019). The work of Del Tredici et al. (2019) showed that an approach that merges information on one's language in their posts and that on their social connections can bring a substantial improvement in user classification task, including hate speech classification.

However, there is a major drawback when it comes to the feasibility and reproducibility of the approach to hate speech detection described by Del Tredici et al. (2019) that has to do with data accessibility. Data regarding the user's activity are often inaccessible as most datasets contain only direct links to social media posts or a single post per user, as it is the case with the dataset proposed by Founta et al. (2018), which does not allow mapping user's network characteristics to so fine a degree like previous research (e.g. Del Tredici et al. (2019), MacAvaney et al. (2019)). Data accessibility was also the problem in this thesis project (see Section 3 for more details) and so the question is how accurate a model that is trained on basic information about the post's author such as the language they use and the list of people they follow on the platform can be. Another question is whether or not it is possible to correctly identify haters when language information is impoverished or even missing; given the nature of the homophily phenomenon (McPherson et al., 2001), it is certain that an analysis of one's social connections might provide some additional information on the user. Del Tredici et al. (2019) investigated how accurate the language model is when social connection information is missing but they did not analyze the social connection model's performance on its own.

The research presented in this thesis addresses these two issues. First, it analyzes whether a simple model

that uses nothing but basic information on who follows whom can distinguish haters from non-haters. Since models based on one's language properties are the primary focus in the field of hate speech identification (Mishra et al., 2018), and since they suffer from problems with data quality and quantity (Kovács et al., 2021; Tsourougianni & Ampazis, 2013), another thing this study checks is whether or not it is possible to correct the prediction of the language model using the prediction of the social connection-based classifier, particularly in situations where the language-based classifier's decision is close to the boundary separating the two classes or where the size of the language data available is small. To verify whether or not such a solution would bring any benefit to the hate speech classification setup, this research is based on an experiment that compares two single-source models: language-based and social connection-based, and then puts them together into one joint decision engine to see if one can improve the performance of the other. The following sections describe how this task was approached and the results of the experiment conducted.

3 Methods

Data preprocessing and model building were performed using a number of scripts. They are available under the following link: https://github.com/milanstanisic/MLT-Thesis-LT2215

3.1 Dataset and Data Extraction

Considering the aforementioned drawbacks of contemporary hate speech classifiers linked to problems with data annotation and given that the aim of this experiment was to construct a model that classifies something as either hate speech or not, it was necessary to find a dataset whose labels matched those needed for the conduction of the study and were defined in a clear and concise way. Eventually, as in the case of the research carried out by Del Tredici et al. (2019), the dataset prepared by Founta et al. (2018) was selected. It is a dataset designated for hate speech-related tasks storing IDs of several thousands of tweets – short entries from the Twitter platform that usually express the author's attitude towards a certain issue (Pereira-Kohatsu et al., 2019) – each of which is labeled with one of four tags: normal, abusive, hateful or spam (Del Tredici et al., 2019; Founta et al., 2018). There were several reasons substantiating this choice, they key one was related to the way the dataset was labeled. As Founta et al. (2018) described, their goal was to create a publicly available dataset whose labels were consistent in what they actually describe and unbiased by human annotators, and so they used crowdsourcing to determine for each tweet which label was selected the most frequently. This way, they reduced bias that could have occurred in labeling should only one annotator label the given data instance and ensured that the quality of annotations was sound.

Following the data selection process of Del Tredici et al. (2019), tweets whose labels were other than hateful or normal were removed from the initial dataset. This was done to tailor the data to the exact goal of the research and to avoid mixing closely related terms (e.g. abusive language and hate speech are not the same) which, as Founta et al. (2018) stated, was commonly encountered among work of their predecessors. Instances with missing data were also removed from the dataset. This resulted in obtaining a dataset that contained a total of 33,422 tweets, 1,947 (5.83%) of which were labeled as hate speech. Compared with other researches (see Pereira-Kohatsu et al. (2019) for more details), this number of tweets seemed sufficient to train a language-based classifier.

In the next step, data were extracted from raw tweet IDs so that all information that was needed for the research – username, tweet contents, and the lists of people followed by the user – was gathered in one place. Most of these data are publicly available for any Twitter user, and to access them, a Twitter account was created specifically for the purpose of data crawling. Data were collected in accordance with Twitter's Terms of Service available under the link https://twitter.com/en/tos which allowed use of users' data by other individuals, and the preprocessing of data was conducted using a series of algorithms to minimize direct human interaction with users' information, resulting in irreversibly encoding all data, including usernames (see Section 3.2 for details). Additionally, data collection process was conducted in accordance with GDPR guidelines concerning personal data use for research purposes. To further protect the privacy of users whose tweets were contained in the original dataset, neither names nor tweet examples are displayed in this paper. Original, unencoded data were permanently destroyed once the research was concluded.

Due to problems encountered with accessing tools available via Twitter API, alternative methods were used. Data were crawled using two Python extensions that were publicly available on GitHub under a GNU license and a MIT license. The first of the extensions – *snscrape* – was developed by JustAnotherArchivist (2023) and served the purpose of collecting basic information on each tweet in the dataset, namely its textual contents and its author. The other extension used was *Scweet* (Jeddi, 2022) and its source code was used to extract names of users followed by each author in the dataset. Source codes of both extensions were modified to satisfy local system requirements and the way data were organized during the research process.

However, due to the slow pace of the latter extension and the fact that for some users data concerning their following lists were not accessible, information on lists of followed people and channels was available for only 399 users. This was a major limitation in this research but nevertheless, it allowed a small-scale analysis of whether or not the solution proposed in this paper works for analyzing a small subset of users and whether or not it could be scalable for larger datasets.

3.2 Preprocessing and Cleansing

To transform raw and scraped contents into a machine-readable form, it was necessary to undertake further steps with regards to data preparation. Using NLTK, a Python extension developed by Bird et al. (2009) for miscellaneous Natural Language Processing tasks, the first step involved encoding raw text from each tweet in the dataset so that it could be fed as a set of features to the model. To extract and quantify information contained within each tweet, a Bag-of-Words approach was used; such approaches assess semantic information on the basis of words that are contained in the linguistic production at hand, disregarding their order (Jurafsky & Martin, 2021). This way of dealing with hate speech classification and detection has already been used in the field (Kovács et al., 2021). To do this, raw tweets were first tokenized and then lemmatized – this process ensured that no word is deemed as multiple independent words due to various inflectional forms it can take in both written and spoken language (Jurafsky & Martin, 2021). This procedure transformed each tweet into a list of lemmatized words, which made these data ready for the next step of preprocessing.

The key stage of data preparation involved transforming tweets and their constituent words into word and document embeddings that could be read by models as features (Jurafsky & Martin, 2021). Since there are multiple ways such embeddings could be generated (Badjatiya et al., 2017; Jurafsky & Martin, 2021), it was necessary to decide which of them would be the most suitable. As Badjatiya et al. (2017) report, classifiers that use embeddings generated through Machine Learning and Deep Learning outperform those generated in a manual way, and so the focus was put on Machine Learning-based methods of computing such embeddings for each word. Eventually, FastText embeddings, originally defined by Bojanowski et al. (2017), were used in this research. This choice was motivated by the fact that such embeddings are resistant to the problem of OOV (Out-Of-Vocabulary) words that could appear in the test set as each word is represented by a sum of its n-gram vectors, which allows computing an embedding even if the word did not originally appear in the training set (Bojanowski et al., 2017). This was a crucial matter in this research since tweets and other kinds of content on social media often contain slang words and abbreviations that are not present in regular corpora (Tsourougianni & Ampazis, 2013). Additionally, FastText embeddings, contrarily to frequency-based embeddings or TF-IDF vectors, are not sparse, which makes computations more efficient (Bojanowski et al., 2017; Jurafsky & Martin, 2021).

Although it is possible to download pre-trained FastText embeddings that are publicly available, this approach was not selected in this research. Instead, custom word embeddings based on the entirety of data preprocessed in earlier steps were generated during the next preparation stage. This was because the pre-trained embeddings that were accessible online were based corpora formed from Wikipedia data whose language characteristics are markedly different from those of data typically encountered in social media (Jurafsky & Martin, 2021). Training embeddings on the data gathered by Founta et al. (2018) allowed them to reflect true word interdependencies encountered on online communication channels, and allowed encoding slang words and abbreviations that would possibly not appear among the pre-trained embeddings. Before training embeddings, all punctuation marks and link fragments that resided after tokenization and lemmatization procedures were removed to further cleanse the data from irrelevant elements. To save computational resources, the length of each word embedding was fixed at 50. Then, to capture general semantics of the whole linguistic production, each tweet was represented as an average embedding of representations of all words that were found in it. Embeddings generated this way were then used as feature vectors for the language-based classifier.

A slightly different approach was taken when data concerning the lists of accounts followed by the given user were preprocessed. In accordance with the homophily principle defined by McPherson et al. (2001), it was inferred that haters would likely be following similar, if not the same, accounts. However, due to the abundance of users of social media, checking these connections across all users would quickly become intractable. Because of that, it was necessary to select a suitable subset of accounts that were followed by a considerable part of the users in the dataset and check if there were any patterns through training a classifier. An algorithm was used to select the most popular accounts. First, it gathered all accounts that appeared in users' following lists, and then for each account, it counted how many users follow it. The next step involved getting rid of less popular accounts by cutting off all those whose number of followers was below 0.999 quantile of the sample, particularly those that were followed by only one user. This procedure reduced the number of accounts taken into consideration in pattern seeking, and it was inspired by the methodology of Del Tredici et al. (2019) who removed solitary authors from the dataset since they did not have any connections with others.

After the initial reduction of the list of all accounts followed down to a list containing accounts with the greatest popularity across the dataset, vectors storing information on social connections were computed from the followed lists of each user. Each account was treated as a token that was encoded as a binary value in a fixed position in the vector. Each value in such a vector was set for 1 when a given user followed the corresponding account and 0 otherwise. Their length equaled the size of the list containing the most popular accounts, with each number corresponding to the respective account in the list. However, considering the collective number of distinct accounts in the list, even after this initial follow list size reduction, it was necessary to reduce the size of the vectors further.

Further reduction was done by computing the mutual information metric between each account encoded in so far obtained vectors and a prediction produced by the social connection-based model. Mutual information is a measure that allows quantifying the degree to which a given variable explains the other – the greater its value, the greater is the degree of relatedness between the two variables (Latham & Roudi, 2009). Computing this measure for each account in the list allowed further selection of only those accounts that seemed to contribute to the social connection-based model's decision to the greatest degree. In other words, accounts whose mutual information score was the highest were more probable to be linked with either haters or non-haters, and thus they provided information that helped distinguish between the two classes of concern. This procedure resulted in producing vectors of zeros and ones that contained information on only those accounts whose mutual information value was above the 0.75 percentile of all mutual information values. For almost all users, with a negligible number of exceptions, those vectors were non-zero, which allowed using them as input data for the social connection-based classifier. This ensured that accounts with this metric equal to zero or relatively low, nearly negligible values were not taken into account by the models.

3.3 Data Splits

The last step of preprocessing involved splitting the data into training and test sets. The blatantly small number of training examples available for the connection-based classifier posed a major risk of model metrics being highly biased and dependent on the contents of each subset. Therefore, the splits were customized to make sure subsequent analyses of the results were inferred on data instances that belonged to the test set in both the language-based classifier and the social connection-based model. Additionally, the experiment was run 30 times so that models' metrics presented would be less biased by a specific split structure, and for each repetition, these splits were different. This subsection describes the exact way data were split at each of the 30 cycles.

Before each cycle, all (399) instances for which the information on lists of followed accounts was available were assigned to the test set for the language-based model, and all other instances became a part of the training set. This operation made it certain that regardless of the way the data would be split into training

and test sets in the social connection-based model, the test data for that model would also be the test data for the language-based model; consistency across samples was the key to conduct a reliable analysis of model performance given the shortage of data that was faced in this study. When it comes to the training set preparation, one issue with the dataset that had to be approached was the class imbalance. The fact that only 5.83% of instances were labeled as hate speech posed a major risk of the language model failing to capture the characteristics of hateful language as they would be overshadowed by the majority class (Daumé III, 2017). To reduce the effect of class imbalance, a random undersampling procedure was carried out. Although Daumé III (2017) suggested that oversampling of the minority class can be used to alleviate this effect, this solution was not used to reduce the risk of overfitting to specific examples or specific authors (Arango et al., 2019). After the procedure, the number of instances that belonged to the negative class (i.e. *normal*) was approximately equal to the number of positive (*hateful*) examples, which made the number of records in the training set of the language-based classifier equal to approximately 3,500.

Undersampling procedure was not performed on the data fed to the social connection-based algorithm since in that dataset excerpt, the class imbalance was not an issue as nearly 25% of 399 records used were labeled as hateful. This proportion ensured that a lack of either positive (i.e. *hateful*) and negative (i.e. *normal*) data instances in one of the sets caused by split's randomness would be very unlikely. When it comes to the split of the data available for the social connection-based classifier, a random split was performed with around 75% of the data being assigned to the training set, and the remainder being assigned to the test set. This means that eventually, there were a total of 100 data instances that the analyses of the results were based on. The differences between proportions of each class in traning sets for each of the two single-source models were present because for the language-based model, an approximate fifty-fifty proportion gave the best results.

3.4 Models

To answer both research questions, a total of three solutions were built, two of which were the single-source models already mentioned: the language-based and the social connection-based classifier. Both of these models were binary classifiers that produced either 0 or 1 as their output value, standing for normal and hateful labels, respectively. Both single-source models classified data independently of each other as they relied on different kinds of information to make their predictions. While the first of them – the language-based classifier – used values stored in generated tweet embeddings, the second one, namely the social connection-based model, relied on features stored in vectors containing simplified and encoded list of followed accounts. These differences between kinds of data were also the reason why each of the single-source algorithms was trained independently of the other.

There is a plethora of Machine Learning algorithms that can be used in classification tasks, and plenty of these classifiers have been used in hate speech detection (Arango et al., 2019; Badjatiya et al., 2017; Goodfellow et al., 2016). Following the results presented by Badjatiya et al. (2017) regarding the performance of various language-based classifiers, it was decided that the most promising results can be obtained using a gradient-boosted decision tree (GBDT). The use of artificial neural networks (ANNs) that are the state-of-art solutions was also considered as such techniques are proven to outperform standard, non-netwirk-based classifiers (Del Tredici et al., 2019). However, given that neural networks need abundant data resources to be trained effectively (Goodfellow et al., 2016), this approach was eventually not selected in this research as there were just a few hundreds of examples the networks could be trained on. The problem of hate speech classification could have also been tackled using Naïve Bayes classifiers which have been proven to be successful in this task (Fatahillah et al., 2017) but eventually they were not used due to some advantages of using decision trees-based methods (see below).

GBDT is an algorithm based on ensembles of decision trees, with a difference that it uses gradient boosting to make each subsequent tree better than the preceding one, which generally makes the model predict

better than random forests, the original algorithm it is based on (Goodfellow et al., 2016; Ye et al., 2009). The selection of these algorithms was justified by their relative transparency as they allow tracing the way the model made its decision explicitly, contrarily to Naïve-Bayes classifiers which are not as explainable (Goodfellow et al., 2016; Ye et al., 2009). Another reason why this classification method was chosen as the baseline one was that for each data point, it gives out its probabilities of belonging to each class, and that information was necessary for the third algorithm, namely the decision engine. For this particular reason and given that the classification goal was identical across both single-source classifiers, the social connection-based model was also a gradient-boosted decision tree. In both cases, the maximal depth of a decision tree equaled 4 as this value brought the most satisfactory results in initial tests.

The third solution, namely the aforementioned decision engine, was essentially composed of the two single-source models. The logic behind it was fairly simple: as each single-source model produced a list of class membership probabilities, it analyzed the list of these probabilities for each model. Then, for each data point, it extracted the highest probability and checked which single-source model it comes from and which class it was assigned to. In other words, the decision engine verified which of the two single-source classifiers showed a greater degree of certainty when making its predictions; the higher the probability observed, the greater certainty. The final decision was the decision of the classifier that produced the highest probability. Such an approach was motivated by the fact that two fairly different kinds of characteristics - language and social connections - had to be combined. This resolved the issue brought by Tsourougianni & Ampazis (2013) that had to do with the tweet length which might produce unreliable results, and helped determine whose prediction was more reliable. Figure 1 illustrates this logic with an example.

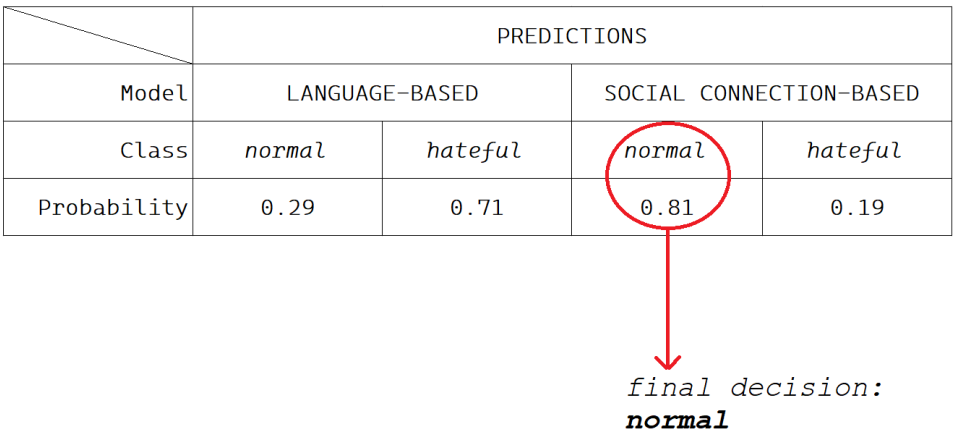

Figure 1: an example of the logic behind the decision engine. Out of all four probability values, the highest one was generated by the social connection-based model. Therefore, the decision engine let the social connection-based model assign the eventual class.

After data preprocessing, a series of the aforementioned 30 experimental cycles was run. Section 4 extensively describes the outcome of the experiment.

4 Results

Due to the aforementioned problems with data scarcity, it was likely that the results would be largely dependent on the way the train-test split was performed. Therefore, to mitigate bias that could have been introduced by the structure of each subset, each of the two single-source models was trained multiple times, with a different data split at each training cycle. Although due to limited data, the test set of the language-based model had to remain identical in each simulation (see Section 3.3 for more details), its training set differed because undersampling at each run was random and resulted in the set containing different data instances. In the case of the social connection-based model, since data were not numerous, both training and test sets differed at each simulation as the same data batch was split randomly. The training-test cycle of both single-source models and the decision engine was performed 30 times, which allowed a better generalization of the results and avoided the potential bias caused by the structure of splits.

Overall, both single-source models struggled with overfitting as their performance metrics were much lower for new data instances. Overfitting occurs when the model relies too much on the training data it has already seen, which makes it predict poorly on unseen instances (Goodfellow et al., 2016; Daumé III, 2017). This was most probably caused by insufficient data resources, an issue already observed by Arango et al. (2019) and Pereira-Kohatsu et al. (2019) which was beyond the control in this experiment due to limitations with data accessibility. Even though data undersampling alleviated this issue and improved general setup performance, overfitting was still apparent to a various extent across the two single-source models. Nevertheless, both of them exhibited some predictive power and each of them was able to classify some fraction of hate speech.

All three setups were compared and evaluated on the basis of four metrics: precision, recall, F1 and ROC AUC scores. Precision stands for the fraction of correct assignments to the target class out of all assignments to that category and recall reflects the fraction of target class instances that were caught by the model (Goodfellow et al., 2016). While these two indicators explicitly show the characteristics of data assignment, the other two metrics - F1 (a tradeoff between precision and recall) and ROC AUC - show how the model generally performs. Out of these two, given limited data resources and overfitting issues encountered, the latter was especially relevant for model evaluation in this study as it allowed determining whether or not the model's decision is random (Goodfellow et al., 2016). The following sub-sections describe the results in-detail and make a comparison across different setups tested.

4.1 Language-Based Classifier

Metric	Average Value - training	Average Value - test
precision	0.81	0.38
recall	0.86	0.88
ROC AUC	0.81	0.7
F1	0.83	0.52

Table 1: Average performance metrics of the language-based classifier.

On average, the language-based classifier was able to catch a vast majority of hate speech both during training and test, which was reflected by relatively good recall scores. However, the precision of the model dropped significantly from 0.81 down to merely 0.38, which means that despite the model's ability to catch nearly all hate speech in the data, it became oversensitive and produced lots of false positives. This entails that the model assigned the majority of data points to a positive class, which is undesirable in terms of the model's behavior; in hate speech classification, the classifier should not label the majority of data points as positive. Considering that the size of the test set the metrics were computed on equaled 100 with around 25 hate speech instances, the values of precision and recall shown in Table 1 entail that approximately

$\frac{recall*n}{precision} = \frac{0.86*25}{0.38} \approx 57$ data instances, thus the majority, would be labeled as positive; this would generate 32 false positives, nearly a third of the whole set examined. Were the model comparably good on the test set as on the training set, it would classify a comparable number of examples but it would produce much fewer false positives. The additional drop by approximately 0.12 of the ROC AUC value also suggested that the model became oversensitive and that its predictions became more random for the test set, which highlighted the problem of overfitting.

4.2 Social Connection-Based Classifier

Metric	Average Value - training	Average Value - test
precision	0.98	0.55
recall	0.59	0.29
ROC AUC	0.8	0.6
F1	0.74	0.37

Table 2: Average performance metrics of the social connection-based classifier.

The issue of overfitting was even more apparent in the social connection-based classifier as merely a third of all hate speech instances was classified correctly, which is visible in Table 2. This was certainly caused by insufficient data resources the model was trained on; a lack of diversity in data made the model learn what connection patterns of haters look like on the basis of too few training examples, which made it not generalize well when new data were encountered. This lack of predictive power is also reflected by low ROC AUC and the remarkable difference in the value of this indicator between the training and the test set, which suggests that the model's strength was not satisfactory and that predictions were much more random for the new data. However, this weak performance of the model did not necessarily mean that its predictions would not be beneficial to a combined setup as it might have been that the classifier produced strong predictions for correctly classified data points. To verify whether or not this was the case, the predictions on the test set were analyzed further.

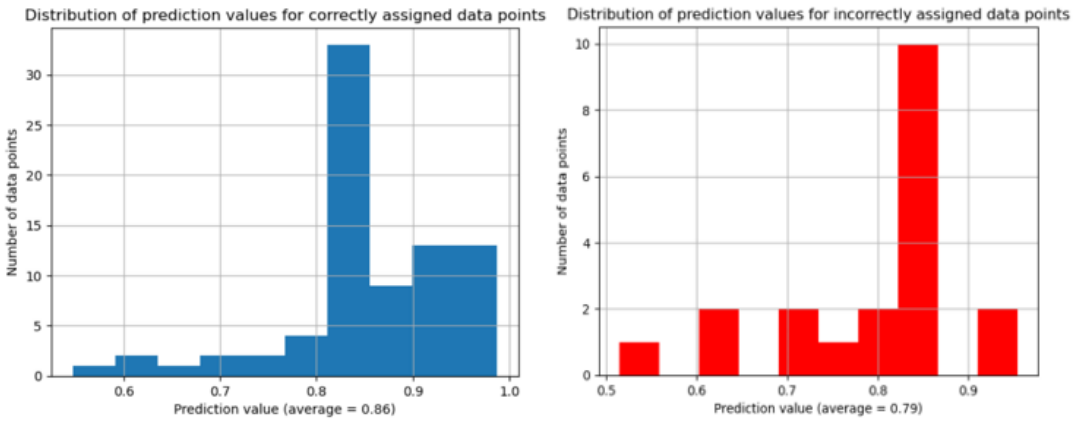

Figure 2: the histograms showing values of class probabilities assigned by the model. The left histogram shows values for the correctly classified data points while the right one shows values for incorrectly classified data points.

Figure 2 shows how probabilities of the classes, i.e. actual model predictions, were distributed across one of the test sets simulated. It quickly becomes noticeable that for correct predictions, the probability of the predicted class appears higher. To verify this, given that these probabilities were not normally distributed, a Kruskal-Wallis test was carried out to see whether or not average prediction values are different between correct and incorrect predictions of the model. The results showed a significant difference between the two samples ($H = 7.92$, $p = 0.005$, df = 1), which, considering that the average prediction value was higher for

the correct predictions, confirms the hypothesis that the model was more certain when it guessed the class membership correctly. This entailed that even though this classifier performed poorly, it could positively impact the performance of the joint solution (i.e. decision engine) where it was correct.

4.3 Decision Engine; Comparison Across Models

Metric	Average Value
precision	0.58
recall	0.53
ROC AUC	0.7
F1	0.55

Table 3: Average performance metrics of the joint decision engine on the test set.

Before running analyses across models, it was verified whether or not samples of each of the metrics analyzed were distributed normally. A series of normality checks was run to verify whether or not normality assumptions are met, which allowed choosing appropriate statistical tests in later stages of the analysis. The tests were carried out using algorithms developed by Virtanen et al. (2020) that are available under Python *scipy* module. The distribution of all metrics in the tests set, regardless of the algorithm, appeared normal (all p-values of normality tests were above 0.05), which allowed applying a series of Student's t-test to compare metrics across models (see appendix A for a table summarizing all metrics obtained in each of the 30 simulations run). All conclusions that follow were inferred with a confidence interval (CI) equal to 0.95.

Since multiple pairwise comparisons were conducted, original p-values were corrected through applying a Bonferroni correction. This correction did not affect the results, all inferences on differences and indifferences remained intact. Table 4 summarizes the results of the series of t-tests described below before and after Bonferroni correction. LB, SCB and DE abbreviations stand for the language-based model, social connection-based model and the decision engine, respectively.

Metric	Models Compared	$t(58)$	p-value	p-value Bonferroni
recall	SCB and DE	-9.149059	<0.0001	<0.0001
recall	LB and DE	14.888286	<0.0001	<0.0001
ROC AUC	LB and DE	-0.105638	0.9162	1.0
ROC AUC	SCB and DE	-6.798296	<0.0001	<0.0001
precision	LB and DE	-8.023940	<0.0001	<0.0001
precision	SCB and DE	-0.939987	0.3511	0.7022
F1	LB and SCB	7.807772	<0.0001	<0.0001
F1	SCB and DE	-7.148629	<0.0001	<0.0001
F1	LB and DE	-0.932093	0.3552	1.0

Table 4: A summary of the results of statistical tests. LB, SCB and DE stand for the language-based model, the social connection-based model, and the decision engine, respectively.

The decision engine that made the final decision on the basis of which model was more certain about its prediction (see Section 3.4 for details) was able to catch around a half of all hate speech instances in the test set, which was reflected by its recall value. This is a significantly better score than that of the social connection-based model ($t(58) = 9.15$, $p < 0.0001$) but also a significantly worse result than the one of the language model ($t(58) = 14.89$, $p < 0.0001$). Its ROC AUC value appeared statistically indifferent from the one of the language model, which was confirmed by a Student's t-test that was run on the two samples of ROC AUC values from all simulations run ($t(58) = -0.11$, $p = 0.91$). While the decision engine's predictive power, expressed through the ROC AUC score, was comparable with the language-based model,

it was significantly better than the one of the social connection-based model ($t(58) = 6.80$, $p < 0.0001$). Additionally, the decision engine managed to produce less false positives than the language-based model. This is visible through a significant increase in precision on the same data excerpt, which was confirmed by yet another Student's t-test ($t(58) = 8.02$, $p < 0.0001$). Precision of the decision engine was statistically indifferent from precision of the social connection-based model alone ($t(58) = -0.94$, $p = 0.35$).

To determine whether or not one solution outperformed another, additional two-sided Student's t-tests were carried out to see if there are any significant differences between their F1 scores. F1 score is a tradeoff between precision and recall that allows assessing general performance of the model (Goodfellow et al., 2016). The results showed significant differences between the performance of the single-source models ($t(58) = -7.81$, $p < 0.0001$), meaning that the language-based classifier generally performed better than the social connection-based classifier, which was most probably linked to the size of the datasets used to train each of the models. The results also showed significant differences in performance between the social connection-based model and the decision engine ($t(58) = 7.15$, $p < 0.0001$), which shows that the decision engine also performed better than that single-source classifier. However, the difference between the performances of the decision engine and the language-based model was statistically insignificant ($t(58) = 0.98$, $p = 0.35$), which entails that these two solutions performed comparably well.

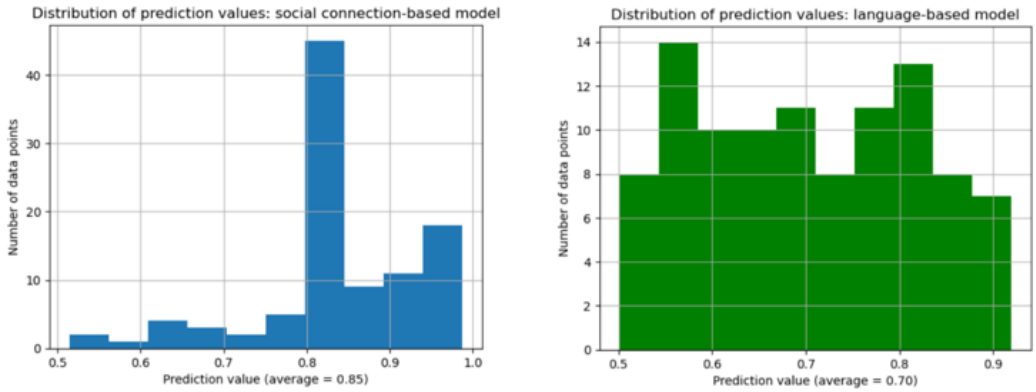

Figure 3: the histograms showing distributions of values of probabilities of the classes assigned by each of the single-source models. These figures were generated from one of the simulations run. Note how the probability mass is shifted towards the highest values in the left histogram compared with the right one.

An additional comparison of probabilities of assigned classes generated by the models underlying the decision engine (see Figure 3), regardless of their correctness, was carried out to determine whether or not one of the models was more certain than the other. Such an analysis allowed determining which of the two models, if any, had a statistically greater influence on the decision engine. Since neither of the two distributions shown in Figure 3 was normal, a Wilcoxon test was carried out to determine whether or not the difference in the average prediction value across the two models was significant. The test conducted revealed a statistically significant difference between the two samples ($W = 505$, $p < 0.0001$). This means that the social connection-based model whose average prediction value was higher assigned classes with statistically greater certainty. This means that the decision engine relied on the social connection-based model even though its performance was poorer with regards to classifying hate speech.

5 Discussion

The results obtained in this study were generally not satisfactory, mostly due to issues with data that were encountered along the way. Their insufficient volume and limitations concerned with their characteristics, such as restricted text input length (Tsourougianni & Ampazis, 2013) or access to nothing but lists of followers that do not reflect actual relations between users and their strengths (Del Tredici et al., 2019), caused the models to perform worse than in studies preceding the one described in this paper. The low values of ROC AUC provide additional insights on the setups' performance as they show that the models' predictive power was not great, and that probably it was difficult for the models to discriminate between the two target classes. This might have been a result of two possible factors, one of which being the possible unrepresentativeness of data with regards to the population; even though the dataset collected by Founta et al. (2018) was carefully gathered through crowdsourcing, it might still have been the case that they included only a subset of all hate speech forms. Another plausible explanation is that they did not have any features that allowed a strong discrimination between the two classes. Nevertheless, although the performance of the models presented in this study was not impressive, their metrics managed to present some noteworthy insights.

This experiment, like many others preceding it, e.g. Mishra et al. (2018) or Del Tredici et al. (2019), showed that semantic content of one's linguistic production is not the only information that can be used to determine whether or not someone is a hater. This was visible through the ROC AUC scores that, although they were not high, were above 0.5 (i.e. the predictions were not random) for all the three solutions tested, solidifying thus findings of previous research on such a possibility. For instance, information concerning social connections of an individual can provide insights on one's group membership – this is thanks to the homophily phenomenon which is also present across online platforms (McPherson et al., 2001; Zook, 2012). Given that accessibility of data resources regarding hate speech and social connections of users can be limited (MacAvaney et al., 2019), this research checked if even the most basic user information, namely their list of followers, can enhance hate speech classification setup in the event of data scarcity. Although both constituents of the joint decision engine overfitted, the proposed decision engine merging their predictions still managed to bring some insights concerning whether or not performance metrics could be improved.

The most intriguing observations can be drawn on false positives i.e. instances where the classifier labels something as the target – hate speech in this particular example – while the given instance is actually not the target (Goodfellow et al., 2016). Section 4 revealed that they were the issue particularly in the language-based classifier as its precision was low. Considering the test size and class proportions, it resulted in the classifier labeling the majority of data points as hate speech. This was undesired since hate speech constitutes the minority class not only in the dataset the models were trained and tested on but also across datasets used in previous research (Founta et al., 2018). Low precision meant that the language-based model overestimated the probability that a given tweet is hateful and that it assigned too many instances to the hate speech class. The low value of this metric was a result of overfitting; while the model performed fairly well on the training set and distinguished tweets between hateful and non-hateful accurately, its performance in terms of precision dropped due to not generalizing well; instead, the model learned the characteristics of training data too finely.

Even though the social connection-based model performed worse on the same data excerpt in terms of the fraction of hate speech it was able to catch (aka recall), it did not produce as many false positives as the language-based classifier. This generally entailed that the model made a decision to classify something as hate speech only if it was strongly convinced that the author's social connections clearly exhibit patterns related to haters. This was confirmed in the distribution of prediction probabilities shown in Figure 3 (the left histogram) which was shifted towards higher values, which meant that hardly ever did the model classify something with the prediction value almost on the edge of the decision boundary (i.e. $P(class) \approx 0.5$). This

fact, although the predictive power of the model was not satisfactory since it did not catch many hate speech instances, implied that once it made a decision about something belonging to the hateful class, this prediction was strong. It was inferred that using the model's predictions might then be useful in correcting the decisions of the language-based classifier which was less certain about its predictions (see the right histogram in Figure 3).

Combining the predictions of these two models and selecting the one with the highest probability score resulted in partially resolving the issue with low precision that was faced by the language-based model as this metric increased to a level comparable with the social connection-based model. Although this caused recall of the joint setup to be lower than the one exhibited by the language-based model, the decision engine produced much fewer false positives while also being able to catch more hate speech instances than the social connection-based model. This can be said to be the improvement in performance that was sought as the problem that comes with overfitting, namely a large drop in precision, was mitigated. Given that the decision engine improved precision of the language-based model and recall of the social connection-based model, the answer to the research question regarding whether or not the proposed decision engine can improve the general setup performance (RQ_2) is **yes**, a solution that combines predictions of multiple models, each of which relies on different information, can mitigate issues linked to overfitting that manifest themselves in poor metrics.

The answer to the first research question concerning the differences between the single-source models, RQ_1, is **yes** since statistical tests that compared average F1 scores of the two single-source models revealed significantly higher performance of the language-based model. However, it is not so straightforward to say that that model was actually better. This is because model's performance can be defined using multiple metrics, each of which explains precisely what the given model is good at and what it has problems with. The first classifier – the language-based one – exhibited substantially better recall, meaning that it was able to catch more hate speech instances in the same dataset. However, as it is mentioned earlier, its precision was low which resulted in unwillingly classifying too many non-hateful tweets as hate speech. The other model – the social connection-based classifier – suffered from the opposite issues; although its precision was better which means that more of its assignments to the hate speech class were correct and that it produced less false positives. Unfortunately, its recall was not satisfactory. This lack of straightforwardness in answering this research question is caused by not only different weaknesses of the two models, but also the fact that they faced major issues with overfitting that severely hampered them in achieving satisfactory performance.

6 Limitations

The primary reason behind the problems with overfitting that, despite the best efforts, were encountered in this research was concerned with insufficient data resources. The single-source models were trained on blatantly scarce data resources composed of merely a few hundreds, and even fewer – just 399 – in the case of the social connection-based model. Machine learning models generally need substantial data resources to generalize well (Daumé III, 2017; Goodfellow et al., 2016), and this could not be provided in this experiment due to technicalities beyond the control of this research. This in turn made the classifiers overfit as they failed to generalize on data; having been trained on an insufficient number of examples, the models captured nothing but characteristics of a small number of instances. When faced with new data whose characteristics reflect a greater degree of diversity, both models failed in being accurate in their predictions since they learned only a few characteristics of haters, their social connections and the language they use.

Had the data resources been larger, the performance of the single-source models would probably have been much better and the overfitting issues would have been less persistent in both of them. Consequently, the decision engine based on these models would have reached much more impressive results; considering that it managed to increase the general performance of the setup even when the models were poor, it can be guessed that it would bring a comparable improvement for models of better quality as well. Another aspect that exacerbated the problem of insufficient data resources was their nature. As aforementioned, tweets, due to their constraints in length and use of slang words and abbreviations to make up for the text volume make their readability and semantical processing difficult and are thus challenging for the models to classify (Tsourougianni & Ampazis, 2013). Additionally, limited information on social connections of people made the social connection-based model quite constrained. But even if the data resources were sufficient, it might still be the case that the model would fail to capture lots of instances of hate speech due to the subjectivity of this notion (MacAvaney et al., 2019). It is not known whether a larger data source would capture all instances of hate speech, as labeling of such terms would inevitably include human bias which would most likely restrict the number of hate speech types down to a few subtypes. Therefore, there are three crucial factors that limited this study and might also impair the results of further studies: data quantity, the amount of information they capture and their labeling.

Yet another limitation of this project lied in the way textual data were preprocessed. Even though the document embeddings were generated on the basis of nothing but the dataset used to ensure that they capture all the characteristics of the online language and that they are more tailored to this type of content, their quality was not actually verified. It might have been the case that the quality of the embeddings generated was actually poor due to limitations of tweet data brought by Tsourougianni & Ampazis (2013). The quality of data after preprocessing was not compared with the quality of data that could use pre-processed embeddings. Such a comparison could have brought essential insights as to which embeddings should have been used to encode data but was omitted because of the reasoning at the time of conducting the experiment. It might have been that the language-based model trained on data encoded using pre-trained word representations would have performed better. Unfortunately, this was not verified and leaves room for improvement.

An additional constraint of this experiment could be identified in the structure and logic behind the joint decision engine that let the model with a greater confidence score (i.e. prediction closer to 1) decide which class a given instance belongs to. This architecture was supported by an observation explained in section 4.2 where correct predictions had greater prediction scores than those incorrect. Although this indeed reduced some issues related to overfitting, it remains unsure whether or not such an approach was entirely correct as incorrect predictions might also have remarkably high confidence scores. Moreover, this architecture did not allow using both information sources simultaneously, which did not allow using social connections as contextual information for language content; this did not conform with the necessity of contextualization mentioned by Dobnik et al. (2022). This could also be said about the single-source models implemented as

they were not compared with other kinds of classifiers, which strengthens the necessity of validating (and possibly improving) the findings presented in this study through a comparison of various solutions.

Nevertheless, these limitations highlighted the problems that previous research struggled with and underlined the necessity of resolving them in the future, and even succeeded to propose a new direction in tackling these challenges. For instance, overfitting that according to Arango et al. (2019) affects nearly all recent work and impoverishes the results of even the most efficient algorithms like those built by Badjatiya et al. (2017), was also apparent in the experiment described here. Therefore, this research confirmed the observations of Arango et al. (2019) regarding the usability of training datasets and models' predictive power when new data are encountered. Apart from that, this project managed to suggest a solution to the problem of overfitting whose effect on the performance can be diminished by combining multiple models that use numerous modalities in hate classification. The solution, even though its logic was simple, worked fairly well and improved general setup performance, at least when compared with the performance metrics of each single-source model, which sheds some light on possible expansions of the research.

7 Ethical Considerations

Before considering what could be done next, it is necessary to highlight that this research, along with all studies preceding it and those that could expand it in the future, raises several ethical issues that need to be taken into account. Hate speech is a touchy issue surrounded by controversial topics, such as white supremacy, homophobia, racism, abortion or religion (Ribeiro et al., 2020; Sellars, 2016), which requires approaching this subject with additional care. Moreover, it is a phenomenon that directly affects its targets and has an undeniably strong influence on the society (MacAvaney et al., 2019; Wich et al., 2021) so it is essential that the matter of this study and solutions presented in this paper be reviewed regarding potential repercussions of applying the proposed algorithms or using general insights brought by this experiment to continue research in this area and/or enhance the presented findings.

The most critical ethical concerns regarding this studies lied in what happens with the data of users whose tweets were analyzed. Since tweets were collected from already existing resources that have been gathered years ago by Founta et al. (2018), users whose content was used to train the models were unaware of the fact that the content they had created was used to build a solution that tries preventing hatred on online platforms. Even though, as mentioned in Section 3.1, the data were collected in accordance with Twitter's Terms of Service and GDPR guidelines concerning data collection for research purposes, this unawareness of users regarding what their data are used for is undoubtedly something that cannot take place if solutions proposed in this study were to be introduced on some online platform. According to Prabhu (2019), people whose data are in use should be informed explicitly as to what happens to the information they give, which is often not the case as these disclaimers are too general. When looking at Twitter's Terms of Service, this social platform informs users that what they make publicly available can be used by third parties or individuals, which, although it allowed this research to be conducted, is a vague explanation as to what actually happens to the information they give. In spite of that, before gathering data to build an actual solution that could bring profit, one should always inform users explicitly about what their information will be used for (Prabhu, 2019).

As far as hater labeling is concerned, another major ethical issue comes with controversies around freedom of expression (MacAvaney et al., 2019). While one might say that haters should not be allowed to express themselves to stop the spread of hate speech, others argue that it might be against the freedom of speech, according to which everyone is allowed to express themselves on any topic they want, and so it is not possible to fully exclude such people from social media (MacAvaney et al., 2019). Ostracizing one from the online world is a radical and controversial solution to hate speech suppression, especially since the algorithms are still not fully accurate. Such an exclusion could be interpreted by somebody as a violation of their freedom of speech and could even lead to suing the party responsible for hate speech detection solutions. This would likely to be the case frequently since the definition of hate speech is subjective, meaning that what is considered an abuse of the freedom of speech by one person might be perceived as an opinion that is not abusive of this right (Founta et al., 2018; MacAvaney et al., 2019). Considering that the models presented in this paper overfitted the data, it is likely that data scarcity had caused them to have identified only specific kinds of hate speech which might not be considered hateful by some people – as already mentioned, defining hate speech is often the matter of subjectivity. To avoid potential unwanted consequences concerning the controversy around hate speech and the freedom of speech, solutions for hate speech detection should be used to prevent other users and algorithms from accessing such content (as suggested by (Badjatiya et al., 2017)) rather than excluding creators of controversial content from the internet.

Even if users whose content was considered hateful by the algorithm would not be excluded from the internet, they might still want to find out why their content was considered as such. However, this is where one of the greatest drawbacks of Machine Learning models, namely limited explicability of their decision-making process, comes in. As highlighted by Prabhu (2019), Artificial Intelligence tools used in automatizing numerous mundane tasks that involve data annotation work like a black box, which means that the transformations of

data are often not clear. Therefore, it often becomes impossible to determine why something was labeled as hateful since it is not known which exact features of a given instance contributed to the model's decision. Considering the abstractness of encoded data used in this research and the way information as encoded, hardly would it be possible to fully explain the model's reasoning. Although metrics such as mutual information can be used to determine which factors contribute to the model's decision to the greatest degree, the exact logic behind the decision remains hard to explain (Latham & Roudi, 2009; Prabhu, 2019). This might raise controversies around model's explicability and should also be taken into account, especially in future research and potential developments and applications of the solutions presented in this study.

8 Future Work

Future developments of the ideas conceptualized in this research are largely dependent on the limitations of this experimentation which prioritizes overcoming all constraints. Since the greatest limitation of this research was concerned with data resources whose volume was insufficient to train robust models, the continuation of this research should undoubtedly follow the direction of repeating this experiment with a greater volume of data, particularly when it comes to those data concerning social connections. Such a re-running of this experiment would most probably alleviate overfitting and lead to more reliable results as the findings of this research only suggest that the solution proposed might indeed improve the metrics of the setup. This hypothesis needs to be verified through a repetition of this research on datasets whose quality is generally better, and the first steps in expanding the ideas and verifying insights brought in this paper should be concentrated around data quality and quantity improvement. Only after overcoming these crucial challenges that seriously affected the results of the research can there be room for further experimentation with various model variants and setup enhancements.

Once problems that constituted major limitations of this research are resolved, next steps could involve testing various model architectures available. Running an experiment similar to the one conducted by Badjatiya et al. (2017) where various architectures were compared with each other could give a broader perspective on what kinds of models actually give the best results when joined with the decision engine. Should the problem with user data accessibility be overcome as well, the research would open up to replacing the single-source models proposed in this paper with more sophisticated architectures and checking how the proposed decision engine would affect their predictions. Data accessibility would not only allow testing various model architectures but it would also pave the way for testing different modalities that could also be used to see what other user information sources could facilitate identification of haters on the internet. Linguistic and social information is not the only source that could shed light on haters' characteristics.

Touching on the issue concerned with the way data were generated, further research could test other ways of data encoding and feeding the single-source models with data generated in a different way. Subsequent research could also verify if pre-trained embeddings outperformed the custom ones that were generated and used in this research; this could actually constitute the first step of such an analysis of the influence of the way data were preprocessed. It might turn out that some ways of data preprocessing are more efficient. To sum everything up, the key direction that work that would follow this experiment should take lies in solidifying and validating the decision engine proposed, as well as mitigating all the issues that this research failed to overcome. Given the promising results obtained with such constraints around, future improvements could bring lots of beneficial insights regarding constructing robust hate speech detection models that would be more resistant to overfitting than state-of-art solutions. In other words, the findings described here open a plethora of possibilities regarding expanding simple, yet efficient ways of reducing the effects of overfitting and inaccuracy in classifying hateful content.

9 Conclusion

Hate speech is a problem that has been becoming more and more widespread on social media platforms. It aims to demean its addressees and has potentially dangerous consequences for the society. To diminish its influence on the internet as well as outside virtual communities, efforts of numerous researchers have resulted in developing multiple solution to automatic hate speech detection. However, many of them have been proven to be prone to issues with data overfitting and a lack of generalizability, which imposed the necessity of finding a solution that would reduce the impact of these pitfalls. The study presented in this thesis described a solution which reduces the impact of overfitting on the eventual hate speech classification results. It checked whether or not it is possible to combine two models, each of which relied on different types of data concerning haters' language and social traits, into one decision engine to alleviate issues faced by many other preceding solutions.

Although the single-source classifiers overfitted the data, this research managed to prove that a joint solution based on class probability comparison across models can select the best prediction, reducing the effects of overfitting that manifest themselves in low performance metrics and large false positive rates that contrast with algorithms' performance during training. The decision engine proposed, considering the performance of the models at hand, managed to alleviate effects directly linked to overfitting and performed relatively well. This suggests that the solution to improving the performance of currently used algorithms might be simpler than one might think and that it does not have to involve a complex setup architecture. This can be roughly compared to one of machine learning maxims that sometimes, a simpler algorithm is better than an overly complex one, and this is what this thesis can be concluded with.

10 Acknowledgements

I want to thank my supervisor – Aleksandrs Berdicevskis – for invaluable feedback on my work and the immense support he gave me over the course of conducting this study. I am very grateful especially for hints he gave me with regards to the logic of the decision engine and all remarks regarding the correctness of statistical analyses. I am also grateful for feedback on the research proposal I received from the second reader - Simon Dobnik - who pointed out aspects of the research that are worth touching on. I also want to thank all the people, including my family, friends, and coworkers, who supported me throughout the entire research process and encouraged me to press on despite difficulties I faced along the way.

References

Arango, A., Pérez, J., & Poblete, B. (2019). Hate speech detection is not as easy as you may think: A closer look at model validation. *Proceedings of the 42th International ACM SIGIR COnference on Research and Development in Information Retrieval (SIGIR '19)*.

Badjatiya, P., Gupta, S., Gupta, M., & Varma, V. (2017). Deep learning for hate speech detection in tweets. *WWW 2017 Companion, April 3-7 2017, Perth, Australia*.

Bird, S., Klein, E., & Loper, E. (2009). Natural language processing with python. https://www.nltk.org/book/.

Bojanowski, P., Grave, E., Joulin, A., & Mokolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 2017(5), 135–146.

Daumé III, H. (2017). *A Course in Machine Learning*. Self-published.

Del Tredici, M., Marcheggiani, D., Schulte im Walde, S., & Fernández, R. (2019). You shall know a user by the company it keeps: Dynamic representations for social media users in nlp. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, (pp. 4707–4717).

Dobnik, S., Cooper, R., Ek, A., Noble, B., Larsson, S., Ilinykh, N., Marev, V., & Somashekarappa, V. (2022). In search of meaning and its representations for computational linguistics. *Proceedings of the 2022 CLASP Conference on (Dis)embodiment*, (pp. 30–44).

Fatahillah, N, R., Suryati, P., & Haryawan, C. (2017). Implementation of naive bayes classifier algorithm on social media (twitter) to the teaching of indonesian hate speech. *2017 International Conference on Sustainable Information Engineering and Technology (SIET)*, (pp. 128–131).

Founta, A, M., Djouvas, C., Chatzakou, D., Leontiadis, I., Blackburn, J., Stringhini, G., Vakali, A., Sirivianos, M., & Kourtellis, N. (2018). Large scale crowdsourcing and characterization of twitter abusive behavior. *Proceedings of the Twelfth International AAAI Conference on Web and Social Media (ICWSM 2018)*, (pp. 491–500).

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.

Govers, J., Feldman, P., Dant, A., & Patros, P. (2023). Down the rabbit hole: Detecting online extremism, radicalization, and politicized hate speech. *ACM Computing Surveys*.

Jeddi, Y, A. (2022). Scweet. https://github.com/Altimis/Scweet.

Jurafsky, D. & Martin, J, H. (2021). *Speech and Language Processing (draft)*. preparation [cited 2023 May 12].

JustAnotherArchivist (2023). snscrape. https://github.com/JustAnotherArchivist/snscrape.

Kang, R., Brown, S., & Kiesler, S. (2013). Why do people seek anonymity on the internet? informing policy and design. *CHI 2013, April 27- May 2, 2013, Paris, France*.

Kovács, G., Alonso, P., & Saini, R. (2021). *Challenges of Hate Speech Detection in Social Media*, 2021.

Latham, P, E. & Roudi, Y. (2009). Mutual information. *Scholarpedia*, 4(1), 1658.

MacAvaney, S., Yao, H, R., Yang, E., Russell, K., Goharian, N., & Frieder, O. (2019). Hate speech detection: Challenges and solutions. *PLoS ONE*, 14(8).

Matalon, Y., Magdaci, O., Almozlino, A., & Yamin, D. (2021). Using sentiment analysis to predict opinion inversion in tweets of political communication. *Scientific Reports*, (2021).

McPherson, M., Smith-Lovin, L., & Cook, J, M. (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27(2001), 415–444.

Mishra, P., Del Tredici, M., Yannakoudakis, H., & Shutova, E. (2018). Abusive language detection with graph convolutional networks. *Proceedings of the 27th International Conference on Computational Linguistics*, (pp. 1088–1098).

Pereira-Kohatsu, J, C., Quijano-Sánchez, L., Liberatore, F., & Camacho-Collados, M. (2019). Detecting and monitoring hate speech in twitter. *Sensors*, 2019(19).

Prabhu, S, P. (2019). Ethical challenges of machine learning and deep learning algorithms. *The Lancet Oncology*, 20(5), 621–622.

Ramos, M., Shao, J., Reis, S, D. S., Anteneodo, C., Andrade Jr, J, S., Havlin, S., & Makse, H, A. (2015). How does public opinion become extreme? *Scientific Reports*, 5(10032).

Ribeiro, M, H., Ottoni, R., West, R., Almeida, V, A. F., & Meira Jr, W. (2020). Auditing radicalization pathways on youtube. *Conference on Fairness, Accountability and Transparency (FAT '20)*.

Sellars, A. (2016). Defining hate speech. *Berkman Klein Center Research Publication*, (2016-20), 16–48.

Tsourougianni, E. & Ampazis, N. (2013). Recommending who to follow on twitter based on tweet contents and social connections. *Social Networking*, 2013(2), 165–173.

van Rijmenam, M. & Schweitzer, J. (2018). How to build responsible ai? lessons for governance from a conversation with tay. *Proposal AOM Specialized Conference*.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., & SciPy 1.0 Contributors (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17, 261–272.

Watts, D, J. (1999). Networks, dynamics, and the small-world phenomenon. *Journal of Sociology*, 105(2), 493–527.

Wich, M., Breitinger, M., Strathern, W., Naimarevic, M., Groh, G., & Pfeffer, J. (2021). Are your friends also haters? identification of hater networks on social media: data paper. *Companion Proceedings of the Web Conference 2021*, (pp. 481–485).

Ye, J., Chow, J., Chen, J., & Zheng, Z. (2009). Stochastic gradient boosted distributed decision trees. . *CIKM'09, November 2-6 2009, Hong Kong*.

Zook, M. (2012). Mapping racist tweets in response to president obama's re-election. *Floating Sheep, 08 November 2012*.

A APPENDIX A: metrics obtained in all experimental trials

This appendix is available under the link https://github.com/milanstanisic/MLT-Thesis-LT2215/blob/main/trials_metrics.xlsx