Was it Serious: Memes as Hate Speech

Research Hypothesis:

If memes can be considered hate speech, how will the law and computer science determine censorship without compromising freedom of expression?

Nehemias Jimenez

JURI 496

4/25/2022

ABSTRACT

My hypothesis is that the law must weigh their options in getting involved with online hate speech, particularly memes. The way I back up my research is through an interdisciplinary approach of law and computer science. I will use the law's understanding of hate speech, with reference to cases that use symbols as the center of hate speech can be later analogized to memes, and cases involving internet speech to lay out the legal theory of hate speech. I also will compare different legal perspectives on whether the law should have a say in the internet world based on current laws in place.

This will be combined with current computer science algorithms used by social media companies to filter out hate speech as well experimental ones. This discipline will also have theories on hate speech and use different approaches to building algorithms. What this research will show is an incongruence between the two disciplines in hate speech theory which exposes a problem of first amendment rights of users of social media platforms and the rights of private internet companies such as Facebook to be free from government interference. An interdisciplinary approach was necessary to understand the full scope of online hate speech as well as weigh all the different aspects related to the law and internet’s relationship. It seems that hate speech is the cross section where both law and computer science is applied when the online domain is the space where it occurs. With both disciplines I feel I have an integrated approach which works best towards the end of stopping online hate speech through memes.

INTRODUCTION

Most of the speech we read in the times we live is not found on a newspaper, letter, or book we read rather it happens on the screens of our phones. And they especially happen in the highly connected world of social media. How does the protection of speech and banning of hate speech that applied to these conventional forms of expression play out in social media? Moreover, how does it work when the very form of expression on social media are not mere words on a screen, such as memes. Herein lies the crux of my research in law and computer science. The law being the primordial protector of free speech and hence punisher of unprotected speech and computer science that is charged with the task of creating algorithms used by social media platforms. When speech is done in its many forms, especially online there must be an integration of both disciplines theories and there must be a solution that best serves the purpose of each discipline. In law there are theories from case law by which one must determine how memes can be hate speech according to hate speech theory. While in computer science one must look at different experimental algorithms based on text and speech. Through these, computer science as a discipline seeks to answer how millions of photos, speech, and memes can be safe from toxic and hateful messages. In using both disciplines combined there will hopefully be a uniform way of addressing what hate speech is and how it can go punished. To what degree does a simple filter on computer science overstep the line of what the law permits it to do as the guardian of free speech. Likewise, how can the law create clear cut understandings of hate speech that are computable for algorithms to enforce. This would create a safer internet, which in turn has become the forum of ideas and debate in this time.

CHAPTER I: LAW

1. Introduction
2. Types of Speech
	1. Free Speech Theory
	2. Hate Speech Requirement Theory in Case Law

# *Beauharnais v. Illinois, 343 U.S. 250 (1952)*

# *Brandenburg v. Ohio, 395 U.S. 444 (1969)*

* + 1. *National Socialist Party of America v. Village of Skokie,* [*432 U.S. 43*](https://supreme.justia.com/cases/federal/us/432/43/) *(1977)*

# *R. A. V. v. St. Paul, 505 U.S. 377 (1992)*

1. Contemporary Issue
	1. Online Hate Speech or Online Art
	2. The Two Sides
		1. Government Intervention
		2. Private Ownership Rights
2. Conclusion
3. Introduction

American pride in the First Amendment exists because free speech is believed to be the most important feature of a liberal democracy. However, throughout the years it seems that discourse can be considered hateful and therefore not protected. With the ever-changing landscape of what is acceptable and what isn’t, and more importantly with the creation of the internet, the law needs a clear way of defining and sanctioning online hate speech. This idea of censorship to sanction online speech that is not protected by the First Amendment is a key issue in law. Particularly, the ways we define online hate speech in light of different forms of expression such as memes needs to be addressed thoroughly. The way the discipline of law helps this endeavor is by laying out the jurisprudence behind free speech. It also contains the changing legal precedents on what constitutes hate speech in cases such as *R. A. V. v. St. Paul, 505 U.S. 377 (1992) National Socialist Party of America v. Village of Skokie,* [*432 U.S. 43*](https://supreme.justia.com/cases/federal/us/432/43/) *(1977) Brandenburg v. Ohio, 395 U.S. 444 (1969) Beauharnais v. Illinois, 343 U.S. 250 (1952)*. While *Elonis v. United States, 575 U.S. \_\_\_ (2015)* provides seminal insight into how forms of art, in this case comedy and poem, can be protected by the First Amendment and when it cannot.

1. Types of Speech
	1. Free Speech Theory

Because hate speech has had many different legal definitions, it is best to first understand what free speech is at its core. In some ways deductively, what free speech is not sets up a good precedent of the kinds of speech hate speech is, unprotected. Freedom of speech in the American Constitution has gone through at least four distinguishable stages of function.[[1]](#footnote-1) The first of these is heavily linked to the independence of America from Great Britain which brought up the first stage of what free speech is for. The first function of free speech was to “establish protection of the people against the government”[[2]](#footnote-2). Stage two can be understood as a shift from protection from tyranny or government opinion to “protect proponents of unpopular views against the majority.”[[3]](#footnote-3) As the flow of ideas in our liberal democracy continued the purpose of free speech being protected shifted towards being concerned with keeping people civil towards other ideas even if most people had made up their minds on what to believe. And so, this last shift of free speech protection’s purpose is girded by the idea of “protection of oppressed and marginalized discourses and their proponents against the hegemonic tendencies of the discourses of the powerful.”[[4]](#footnote-4) All these primordial functions of the same amendment deal with power dynamics of people or ideology. So, it seems that this kind of tension-filled activity of “free speech” should be well justified. Michael Rosenfeld delineates four philosophical justifications for free speech protection.

The first of philosophical justification for free speech is what Michel Rosenfeld calls “justification from democracy”.[[5]](#footnote-5) This kind of justification means that without being able to spread ideas and converse about them, the job of governing oneself, as is the purpose when in a democracy, cannot be fulfilled. The next, “justification from social contract”[[6]](#footnote-6) reveals how without free speech, the business, cultural, and even day to day agreements we participate in are virtually impossible. In this sense these first two justifications are alike. The third justification, “justification from the pursuit of truth”.[[7]](#footnote-7) The premise is that truth is incremental, and we move towards a more holistic understanding of issues, even sciences over time. Without free speech, we will never have trial and error in our debates that lead to an incrementally greater amount of true knowledge. The final one, like pursuing truth, is the “justification from individual autonomy”.[[8]](#footnote-8) This justification says that “individual autonomy and respect require protection of virtually unconstrained self-expression.”[[9]](#footnote-9)

This has particular relevance because for any kind of speech, online or in person, to be considered hate speech it must not be interfering with one or more of these justifications and functions. What we will see is how these theories underpin legal decisions and what the US courts carved out as the requirement to constitute hate speech.

* 1. Hate Speech Requirement Theory in Case Law

# *Beauharnais v. Illinois, 343 U.S. 250 (1952)*

One of the first cases that deviated from a prior standard that was set for what constituted hate speech was *Beauharnais v. Illinois, 343 U.S. 250 (1952)*. It deviated from the Chaplinsky framework which was a hate speech standard set in Chaplinsky v. New Hampshire, 315 U.S. 568[[10]](#footnote-10). The standard included “libelous or fighting words”.[[11]](#footnote-11)In 1952, a white supremacist was brought before the Supreme Court for distributing pamphlets saying defamatory things about black people[[12]](#footnote-12). Some of these things included that they were the aggressors of rapes, they were the people to blame for drug abuse and robberies as well. At the time, the current attack was that the man spoke in a way that caused “group defamation”. This was taken out of the idea of individual defamation being outside of the scope of the First Amendment’s protection.[[13]](#footnote-13) In online speech, possibly in memes, messages can be towards or concerning an individual or a group. Both are relevant and for that reason this case fits one part of the bill. In the decision by the Supreme Court, the court found that he had not posed a “clear and present danger “to the community for it to be considered fighting words according to the Chaplinsky framework. The reason why the Court found it was not libel was partly because he had not tried to publish or push this pamphlet with “good motives and for justifiable ends.”[[14]](#footnote-14)Had he tried to claim these as true, such as in a scenario where he was a teacher making students read this pamphlet as a historical and peer reviewed article, it would have been violating Illinois libel law. Or if he had claimed these true towards an end to justify racism it would also be defamation. This matters because here we see intent matters in considering speech as hate speech, and it seemingly must prompt some violence. Moreover, the court argued that it would create conversation that was public in nature. In the eyes of the court, this kind of speech was justified in pursuit of truth. Although controversial the Court deemed that this kind of speech could start a dialogue that would move to the retrieval of knowledge.

# *Brandenburg v. Ohio, 395 U.S. 444 (1969)*

Later in *Brandenburg v. Ohio, 395 U.S. 444 (1969)* a court decision was reversed by the US Supreme Court concerning a KKK leader and group that were on live television saying derogatory things about Jews and Blacks. To add on to these things being said, they claimed that they would work towards accomplishing this goal if the government failed to do so. The Court reasoned that the KKK ``may have advocated violence but had not incited it.”[[15]](#footnote-15) By this they meant that to act out of their political view on people going back to a place of origin as a means of following up their protest was an extension of their First Amendment right. Justice James Douglas made an analogy to a prayer accompanied by a posture and movement of hands[[16]](#footnote-16). In doing so he cements that inciting violence is an imminent threat to hurt an individual or group. This kind of speech was mostly justified under the justification of individual autonomy. Although all justifications can intertwine, it is worthy to note that respect for these people who identify with the group have a right to express those views and act on them in ways that are protected under the First Amendment. This matters because all messages that convey ideas can be interpreted as hateful when it is just a speaker’s autonomous self-expression. This is difficult because it means that there are vile things that must be spoken of from an objective point to understand the belief. On the other hand, people who hold that very same belief have the autonomy to self-expression, so long as it is not directed towards a group or individual to incite violence. And in accordance the legal reasoning is that there can be an advocacy to act on it. How different advocacy is to inciting is difficult. In this case there was no acting and so there is no correlation a court could have made to say it incited a particular act.

* + 1. *Village of Skokie v Nat’l Socialist Party of America*

In this case the kind of incitement to violence was first met because the *National Socialist Party of America v. Village of Skokie,* [*432 U.S. 43*](https://supreme.justia.com/cases/federal/us/432/43/) *(1977)* wanted to walk through Jewish communities.[[17]](#footnote-17) The Court reversed this later because in basic terms, although it was abhorrent to Jews and especially those who were affected by Nazi ideology, it did not grant grounds to stop the expression of ideas.[[18]](#footnote-18) So although a message may have been offensive, it did not mean that it was directed as a hate speech, rather a political message. Interestingly it seems that the only connection to the justifications is that the people in the area of the intended march would have made the Neo-Nazi’s a minority who at the time had ideas that weren’t popular. Rosenfeld finds that in this situation and in the time, the protection of this free speech was like that of the third stage. Some views aren’t popular, and memes and internet speech in general must be granted the right to express those views, even if it is considered outdated or offensive to others. This case does well in expressing that as times change, we believe some things become unacceptable, however the unchanging justifications of free speech theory allow for these thoughts to run in circulation. This in some ways is like allowing some misinformation to run because the truth always wins. Online organizations and websites of less popular alt right views are consistently left on the web to give less attention and advertisement to the ludicrousness that they say.

# *R. A. V. v. St. Paul, 505 U.S. 377 (1992)*

The last key case in history that holds importance in the hate speech requirement is *R.A.V v. St. Paul.* This case in particular takes the symbol of a burning cross in a black family’s front yard. The Supreme Court ended up reversing the initial verdict saying once again that it did not reach the “inciting violence” standard. Symbols in political speech are protected under the First Amendment. According to the court the initial decision was “unconstitutional because it imposes special prohibitions on those speakers who express views on the disfavored subjects of "race, color, creed, religion, or gender.”[[19]](#footnote-19)

 This case is particularly interesting regarding meme because there seems to be a reasoning present that signs or symbols that have a meaning aren’t inherently hateful. Memes in some ways are just symbolic representations of cultural beliefs. The wording themselves is not meaningful, rather the picture gives it its humor or distaste. This puts into question how memes themselves as symbols can be considered hateful. The symbol only serves a representation of thought and nothing more. The reasoning in the case might be analogically used in a legal case concerning a meme. The other case laws chosen are also important when looking at memes posted online possibly constituting hate speech and needing censorship. In *Beauharnais v. Illinois, 343 U.S. 250 (1952)*, the idea that hate speech is based on the intent of the speech is important to us. Particularly, an idea that there is clear and present danger is important because when looking at memes, it is mostly understood as always being an art form of comedy. Can someone point to a meme and say there is danger? In the case *Brandenburg v. Ohio, 395 U.S. 444 (1969)* we see that advocating violence, meaning even memes that seemingly are the darker side may be saying something that can advocate but not incite violence. *National Socialist Party of America v. Village of Skokie,* [*432 U.S. 43*](https://supreme.justia.com/cases/federal/us/432/43/) *(1977)* and *R.A.V* makes the notion that symbols are also a way of expressing ideas. To look at memes as a symbolic expression of some personal or cultural belief would be of importance here in determining its status as potential hate speech and subject to sanction since in *R.A.V* we see it unsanctioned.

III. Contemporary Issue

1. Online Hate Speech or Online Art

The seminal case that combines the reasoning of the previous case into a real online hate speech case is *Elonis v. United States, 575 U.S. \_\_\_ (2015).* This case takes the Facebook posts of a man that were considered threats towards all those cryptically directed to. Some specific phrases he used were from a comedy special that he twisted to make a threat towards his wife. Later using poems, he would do the same to a kindergarten class. The case is important because it is a hate speech to hate crime case that is based on online speech. Moreover, the actual packaging of the alleged threat was in art form. Memes as comedy falls into a kind of art form as well. The Court concluded that there was no credible threat although if the words were packaged outside of prose form it could have been judged differently. A main argument used to reverse the decision was that “a guilty mind is “a necessary element in the indictment and proof of every crime”[[20]](#footnote-20). Although the First Amendment was not checked here, it seems that the intent of the artist who puts out the art can outweigh the possible threat that can be perceived by others. To prove that it was not a threat, the defendant proved he had no intent. What this means is that so far there is little legal evidence that memes can constitute hate speech, although it seems that if they could, it would have to be proven that the incitement to violence, or a credible threat is intended by the speaker.

1. Two Possible Government Reactions
2. Government Neutrality

There are two main lines of thought pertaining to how the government should react regardless of if hate speech is seen in online memes. We shall stick with the premise that it cannot be hate speech because of case evidence. The first of these is the government ensuring government neutrality in all online servers that allow users to post. This would essentially have to change *47 U.S. Code § 230*. Currently, “the provision provides the foundation for the Internet as we know it today.” [[21]](#footnote-21) The legislative change to assure that no memes are banned when they cannot be considered hate speech is something that would possibly be covered by Senator Hawley’s Political Neutrality Requirement.[[22]](#footnote-22) It would also have to change the internet from being considered a public utility. A public utility being “any organization which provides services to the public, although it may be privately owned. Public utilities include electric, gas, telephone, water, and television cable systems, as well as streetcar and bus lines. Public utilities are allowed certain monopoly rights because of the practical need to service entire geographic areas with one system, but they are regulated by state, county, and city public utility commissions under state laws.”[[23]](#footnote-23) This would be tricky only because of what Lily Coad says. She believes that algorithms made by private companies are a type of free speech. She believes that “they express to users, and to the entire world, that the company maintains a certain viewpoint about an idea or topic-or even a single word. As such, those algorithms are protected speech, guarded from government intervention and censorship”. [[24]](#footnote-24)

1. Private Owner Right

The other view is that if memes are not hate speech there is no harm in letting these same companies hold the right to banish and censor what they personally please. If it is true, that these algorithms are a type of speech by companies such as Facebook or Twitter, then it follows to say that it is to stop their constitutional right. Much like an individual who needs autonomy to hear conversations, these private companies would be able to keep and disregard particular speech that their platform does not condone. The difficulty in this path is the global power of the internet. Different people may be censored if a company that provides a Facebook-like service is from Russia or from another country. Or in the opposite case, that which we do not censor is offensive and illegal elsewhere.

IV. Conclusion

Based on the legal theory and case studies, memes have yet to constitute hate speech. However, there is a clear path as to how it might be defined in a court setting. In the current way that things stand, it seems there are two ways the government can assure that hate speech gets sanctioned correctly when it does come to be. Neutrality allows for the uniformity of punishment from algorithms across the internet by overstepping the current legislation that gives power to the providers. On the contrary, to keep it the same where the private sector chooses what happens may hinder First Amendment rights and have international issues.

Chapter II: Computer Science

I. Introduction

II. Theories of Online Speech

1. Hate Speech
2. Exception to Hate Speech Rules

III. Algorithm Building Blocks

1. Sentiment Analysis
2. BERT and OCR

2. Image Captioning

IV. Pros and Cons

 A. Risks and Limitations

 B. Multimodal Solution

V. Conclusion

I. Introduction

Online speech is the platform in today's world that presents the largest audience for any person to be heard. However, its unprecedented growth from a place of mundane conversation to a place with serious political and ideological dialogue has left serious questions about censorship and free speech. The way computer science algorithms promote particular speech and ban others is a conversation that has yet to be fully explored. Without question computer science is the single most important field in the development of technology that accomplishes the forementioned goals. Studying computer science allows an understanding of how it tackle the problem of hate speech being perpetuated through memes. Looking at the technological advancements and combining of different algorithm building blocks, one can look at the best possible algorithms to detect and sanction hate speech.

II. Theories of Online Speech

1. Hate Speech

 Online speech is governed by computer science and the platforms that employ its theories and software. Hate speech’s detection and punishment is also then upheld by the theories in computer science about what constitutes hate speech and what doesn’t. One of the most comprehensive definitions in computer science is that “Hate Speech is a direct attack on people based on race, ethnicity, national origin, religious affiliation, sexual orientation, sex, gender, and serious disease or disability[[25]](#footnote-25).” The definition has some ambiguity such as the words direct attack. The way most algorithm understand this criterion is by whatever standard is imputed into it by humans. Some have also tried to clear up misconceptions about hate speech by classifying what it is not. The first of these misconceptions is that “an endorsement of the organization does not constitute a verbal attack on another group”[[26]](#footnote-26). Meaning that social media algorithms cannot stop a post about how great Trump was as a President, or how abortion is moral even if you find either appalling. The belief that endorsement does not equal a verbal attack is one that computer science imputes into algorithms. The algorithm in turn has no incentive to punish endorsements because they are not direct attacks, rather are the expressions of personal ideas and beliefs. Computer science attempts to create algorithm knowledge that out other kinds of posts that would fall under the direct attack criterion. One of these specific posts would be that with the “unnecessary labeling of an individual as belonging to a group” [[27]](#footnote-27). This is because this labeling usually couples with it particular stereotypes and beliefs that might be demeaning to a said group. An example of this would be “what could you expect from X? They are a (insert group label) after all.” And so, computer science understands that “hate speech often employs well known stereotypes to disparage an individual or group” [[28]](#footnote-28). What is important to keep in mind is that computer science as a field also believes that “while disparaging terms and racial epithets when used with the intent to harm always constitute hateful language, there are some contexts in which such terms are acceptable.”[[29]](#footnote-29)

1. Exception to Hate Speech Rules

Aside from more intuitive exemptions to the rule such as educational texts and publications including words such as the N word in a historical context, group solidarity is also exempted. Group solidarity is a phenomenon also known as word reclamation. It “is a common defence mechanism…. when a group of people reclaim a word and strip it of its hateful nature.”[[30]](#footnote-30) This is evidently seen by the African American community who has “taken back” the N word to show group solidarity. It is now considered a term of endearment to some in the community. This is important for computer science because in almost any algorithm “if the identity of the speaker cannot be ascertained, and if no orthographic or other contextual cues are present, such terms are categorized as hateful”[[31]](#footnote-31). This shows that computer science algorithms try to take into account some of the linguistic habits of our societies. In a basic sense computer science acknowledges that “Sometimes such words are used by a speaker who belongs to the targeted group, and these may be hard to classify without that knowledge”[[32]](#footnote-32). At first glance it may seem like a small exemption, but it has big implications to the way algorithms are designed. And what it also does is set up some basic understandings of what content in hate speech, even in memes, needs to be looked for. From a theoretical perspective, these theories leave “numerous issues involved in defining what constitutes hate speech, which need to be resolved in order to annotate a corpus and develop a consistent language model”[[33]](#footnote-33). Nonetheless they serve as a starting point to the basic building blocks of algorithms to detect hate speech.

III. Algorithm Buildings Blocks

A. Sentiment Analysis

1. OCR and BERT

Sentiment analysis is the analysis done by the computer algorithm that is usually composed of OCR and BERT. It is the process by which an algorithm interprets or categorizes what is on a screen. Related to online speech and memes, direct attacks are not always so obvious. Something as simple as “comments…obscured with an intentional misspelling, presumably to evade a filter employed by the site.” [[34]](#footnote-34) can mess up a too simple analysis. Here in lie two extremely important pieces of software used in building a stronger algorithm, especially for memes. Optical Character Recognition (OCR) is a tool used to take visual images of texts and turn them into machine texts. This is what software such as Adobe accomplish when a pdf is scanned and uploaded whilst the ability to edit as if it were a digital document remains intact. This kind of software works in conjunction with BERT or a BERT-like processor. Bidirectional Encoder Representations from Transformers (BERT) serves as judge in deciding which words are hateful in a text. BERT works for all kinds of text which can either come from OCR or not. BERT translates words into numbers that the machine can read. This is one of the first things that makes Bert different from other engines. It does not just search for words in a text, because specific lettering might be changed for hate speech. This also helps in the normal and very repetitive scenario where famous words of hate speech are not being used hatefully. To overcome the shortcomings of other models, BERT actually looks for the meaning of the word in the context of the sentence. BERT understands for example that “words have a stereotype sense, in that they either antisemitic or not, and we can learn the sense of all words in the corpus from the paragraph labels.”[[35]](#footnote-35). This can be applied to all kinds of stereotype words. And so, all these words are then used in all contexts, and BERT is taught to notice the surroundings words that would weaponize said word. A clear example is that BERT sees a misspelled stereotype word being preceded by the word “stupid” in a post. BERT would automatically assume that the word after stupid would be used hatefully. BERT understands that the context of stupid isn’t in a sentence with good which could be read as “stupidly good”. Rather when seeing a misspelled word such “B1t(h”, it understands the context of stupid is harmful. In contrast something such as “Bitch please” is a way someone might reference someone they are extremely familiar with to say, “stop joking with me”. When BERT is used, with or without OCR, computer science calls sentiment analysis. When it comes to this disambiguation task of words “creating a language model for each stereotype is a necessary prerequisite for building a model for all hate speech.”[[36]](#footnote-36)

 2. Image Captioning

Image captioning is a simple tool gives a word definition to a specific online image. All the images that one finds through a Google search is directly linked with image captioning text each picture has when published. When someone writes the specific description of a photo it matches the image caption and pulls up that photo in the search results. It is also very beneficial with memes since memes always have a picture in it. When dealing with memes “the model is fed an image and outputs a description of that image”[[37]](#footnote-37). Each of these models might be able to get their respective job done, however each on their own have massive gaps in detecting memes. These singular models are what we call unimodal models. They do well on their own but when each has been put to the test of detecting hate speech in memes, they have fallen short in many ways.

IV. Pros and Cons

 A. Unimodal Limitations

Unimodal models are models that take up the challenge of memes. They are unimodal because they are algorithms based on either image captioning, or a BERT-like software. Just the task of finding hate speech in either category will reveal some weaknesses in a unimodal algorithm. These weaknesses and limitations become even more evident when the task of these programs is focused on detecting hate speech in memes. The issues memes present to a unimodal model are what I call the lingual problem, meme nature problem, and the bias problem. The first of the lingual problems is that some passages of media might sound like hate speech until seen in its full context. In other words, some posts that can be quoting some specific source that might not be known by BERT. “These comments are short, with an average of length of 31 words, and lacked the contextual setting in which they were originally found”[[38]](#footnote-38) . An example can be a passage of a fictional book with characters from times past who invoke ideologies. This criterion of not enough words being placed for BERT to understand the words context is more severe with memes. When looking at memes, the number of words might just a mere word or a singular word. Simply put, the words might be flagged because there is no context or not enough context to judge what is happening. The second problem is the meme nature problem which connects how the very nature of memes being a message of both picture and words gets the fullest picture. If one looks at the data from image captioning “it is hard because a picture can represent anything, and it is impossible to compile a dataset that contains every possible object or person.” [[39]](#footnote-39) And looking at words alone can lead to what computer science calls benign confounders. These memes appear to have no normal stereotype words but when coupled with the picture give the fullest meaning. An example is the words “ah, the dish washer is broken again” with the picture of a woman. It becomes inherently clear that “when viewing a meme, a human would not think about the words and picture independently; but understand the combined meaning.”[[40]](#footnote-40) Obviously “mocking hate crime is also considered hate speech”[[41]](#footnote-41) because an attack can be “violent or dehumanizing (comparing people to non-human things, e.g., animals)”[[42]](#footnote-42) or in this case a dishwasher. This proves one of the claims done in a study that “visual cues are much more important than linguistic ones when detecting hate speech memes” [[43]](#footnote-43). Memes have also taken new forms such as the distorted meme in contrast to the classic and modern meme. “Classic or modern memes refer basically to the format and placement of the text.”[[44]](#footnote-44) However when distorted memes come into play with highly compressed and distorted images “it may affect the quality of the OCR recognition and, therefore, the language encoding.”[[45]](#footnote-45). The final critique is the bias critique which has to do with the very datasets that train algorithms and teach them measures of detecting hate speech. The people chosen to do so may not be varied enough to give good criteria to engines like BERT. New and clear “findings show that existing approaches to toxic language detection have racial biases, and that text alone does not determine offensiveness. “ [[46]](#footnote-46) Particularly in the African American community, people chosen to train programs such as BERT misinterpret messages as offensive because of racial bias. There is also a higher “differences in rates of false positives between AAE and white-aligned dialect groups”[[47]](#footnote-47). It also worth noting that when the racial bias creates the dataset for algorithms there are specific linguistic habits from different cultures are missed. “For example, the use of ‘‘419’’ to mean an unwholesome behavior is commonplace in Nigeria”.[[48]](#footnote-48) This is more of a social critique but nonetheless important because if not addressed is “risks further suppressing already marginalized voices” [[49]](#footnote-49).

B. Multimodal Solution

Here in lies computer science’s latest response and answer to the problem of meme hate speech detection. “Multimodal models hate speech detection is a vision and language task”[[50]](#footnote-50) . It tries to combine both approaches together to solve the code of getting the full meaning of memes. Although it seems simple the “hybrid structure makes processing and extracting information difficult since it adds another layer of complexity to the data.” [[51]](#footnote-51) Basically the “system expects an internet meme input and produces a hate score as an output.”[[52]](#footnote-52) Firstly, the OCR begins to translate the pictures text to another kind of text that is readable to BERT or an alike program. It at the same time uses Image captioning to get an accurate description of the meme photo. When compared, like puzzle the pieces fit together. In the example of before of a meme that says “oh, the dishwasher is broken again” with a picture of a woman, the image captioning of a women smiling with the text of BERT would produce a hateful result if BERT has been taught to recognize dishwasher as a demeaning term for a woman. In the many recent studies multimodal models “obtains the best results, followed by vision only case.[[53]](#footnote-53)”. Meaning that in comparison the unimodal model of language only was the worst in comparison during experiments. Most people in computer science conclude that this approach helps beat out the confounders.

V. Conclusion

Overall multimodal models for detecting hate speech work best in comparison to models available. However, “it would still require a human moderator for many of them.”[[54]](#footnote-54) Based on data these multimodal models still preform less than humans 64.73% compared to 84.7% accuracy. [[55]](#footnote-55) Mostly because this model still struggles against distorted memes and the previously mentioned social critique from bias of algorithm trainers. As of now the problem of detecting hate speech in memes stands in computer science.

CHAPTER 3: INTEGRATION

1. Common Goal
2. Hate Speech Definition
	1. Likeness
	2. Differences
3. Conflict in Boundaries
4. Possible Synthesis
5. Common Goal

In looking at both disciplines the goal is to look at the similarities in differences in approach and in theories to get the best integrative solution to hate speech being present in memes. The first and most important similarity in both disciplines is the common goal that they each have in attacking hate speech, especially online. It seems that case law, particularly *Elonis v. United States, 575 U.S. \_\_\_ (2015)* shows that the law has accepted the task of answering what online hate speech is and condemning it in the reasonable manner when necessary. At the same time this opens up the discipline of law as self-aware that memes may enter as the centerpiece of speech in debate. Since it is a form of online speech computer science as a discipline has also felt a need, although not necessarily in conjunction with the law field, to also make effective algorithms to detect and ban hate speech in memes. This separation which will be spoken about in more detail later is because of *47 U.S. Code § 230* setting up computer science to working with private companies and service providers to enforce their codes of conduct regarding speech. Both then have a sense of obligation that sets them on the right foot for integration.

1. Hate Speech Understandings
	1. Likeness

The shared space between both fields also extends into understandings of hate speech on a theoretical level. Both understand that “there are numerous issues involved in defining what constitutes hate speech, which need to be resolved in order to annotate a corpus and develop a consistent language model”.[[56]](#footnote-56) This idea first expressed in computer science writings is upheld in law practice by the consistent looking back on the history of hate speech cases to reason through newer cases. In doing so there is a unity in understandings that hate speech is hard. It is not something that can be done in one sitting but is a process of learning. This is particularly evident as the passage of time sees new words or expressions becoming hateful. The law itself has changed precedent and it is always up for debate at the highest level of jurisprudence what specifically constitutes hate law even when there are standing precedents. Consequently, the computer science field upholds this belief in practice by employing the method of “teaching algorithms”. It denotes a kind progressive learning that is not just sharpened on the technological plane but also in the theoretical categories of what hate speech can look like. The recent research previously referenced from computer science shows how memes have become a mode of speech that needs to be taught to algorithms. Another likeness in the two fields’ understanding of hate speech is that it is not always clear and loud. At times it is very hidden. Computer science puts it that “comments…obscured with an intentional misspelling, presumably to evade a filter employed by the site.”[[57]](#footnote-57) The law also understands this subtly in hate speech. Most law cases would not get before the Supreme Court unless there was uncertainty about the nature of the speech. I believe that *R.A.V v. St. Paul* perfectly exemplifies this. A burning cross was vague and subtle to convey an idea that was brought before a case to determine whether it was hate speech or not. Memes by their nature are also cryptic and symbolic which is why some cannot understand it without the inside knowledge. And both disciplines having this alike understanding gives them much effectiveness in sorting out what is hate speech and what is not. The final shared understanding is that “an endorsement of the organization does not constitute a verbal attack on another group”[[58]](#footnote-58). The case *National Socialist Party of America v. Village of Skokie,* [*432 U.S. 43*](https://supreme.justia.com/cases/federal/us/432/43/) *(1977)*  shows how the law also shares the understanding that not every speech someone is offended with is hate speech. Had that been the case, the National Socialist Party of America would have lost their case. All this goes to showing that both hold the belief there is an intrinsic nature to hate speech that is objective and not just to the whim of the receiver. This intrinsic nature seems to be where their respective understandings diverge.

* 1. Differences

One of the differences in understanding between the fields is based on the fact that computer science algorithms are just a tool used by private companies. What this means it that the standard of what this intrinsic nature of hate speech is detailed by a Facebook or Twitter. And even if they wanted to be independent, creators of the algorithms, when separate from the law, create their own standards and definitions of hate speech used to detect and sanction online hate speech. One such group of algorithm makers from the computer science discipline that was mentioned said that “We define attack as violent or dehumanizing (comparing people to nonhuman things, e.g. animals) speech, statements of inferiority, and calls for exclusion or segregation. Mocking hate crime is also considered hate speech”[[59]](#footnote-59). The emphasis is clearly on the “We”. This is different from the Chaplinsky framework and the building off that precedent. Meaning that in a case of online speech where some kind of mocking was present, there would be two different judgements from the respective fields. One judgement would be arrived by the law saying that inciting violence is not met. And if it is not met, it can’t be hate speech. The law would get to that conclusion by pointing out that there is no present danger in mocking another person perhaps. On the other hand, computer science people who impute the values sanctioned by algorithms might have a different take. The situation only gets more complicated with other private companies with allegiances or ideas in mind of promoting. This difference is not an entirely subjective definition, rather it is just multiple legal theories by the computer science field (or companies that hire them) on what unprotected speech is. They are doing a part of the work the law has already done for them. And this fact might be known by people in both fields. And certainly, there are those who might not like working together as mentioned in the earlier quotes from Lily A. Coad, Compelling Code: A First Amendment Argument Against Requiring Political Neutrality in Online Moderation, (2021). Other such as Benjamin H. Winters who wrote A Clear and Present Danger: The Need for Regulated Accountability for Online Service Providers to Preserve and Promote Free Speech, Notice and Due Process., (2018) run into the greatest conflict to full integration.

1. Conflict in Boundaries

This entire conflict of why the law and computer science don’t want to work together, and more importantly can’t, is *47 U.S. Code § 230.* This creates a conflict of boundaries and forces each to work on their own in attacking hate speech. The law states in a summarized sense that the whole internet and the services within (Facebook, Twitter, YouTube) are private and cannot be infringed on by the government. This is in opposition to the internet being a public utility such as gas or water that while privately owned are government regulated. The conflict then is that when these algorithms ban a meme or speech as hateful, but do not line up to the legal definitions of hate speech, we get an infringement of the First Amendment. The person can take this to the legal world, but at the end cannot get unbanned perhaps because it is the right of the private company. So basically, we have people in both fields saying that the legal world can’t infringe the right of the private company to do what it wants, whilst the legal world wants to do its job in protecting free speech without breaking the other law which says the can’t be involved in regulating the Internet. One big reason why is because there is a fear that the government will have too much control on what is circulating on the internet. Regulating the Internet can indirectly mean regulating the newspaper, educational information and private information of users on websites from the US and other countries. One cannot just create a separate Facebook for every country so that private information of users in Russia aren’t given to the US government even if the government would like the idea of having foreign intelligence. And even if one could, the fact of the matter is individuals have a right to privacy as well. The internet slides down the slippery slope of being public as is the real world whilst also being private like one speaking in their house.

1. Possible Synthesis

The solution to this problem is difficult because of these boundaries. Here I have but one solution. My solution is based on *47 U.S. Code § 230* where the law is changed to account for the conflict of interest, and it is renegotiated what rights of regulation the government might have on the internet. This would allow possibly for a uniformity of what constitutes hate speech and computer science would work with the law, according to the laws precedent to stop hate speech in memes. However, this renegotiation must be solely on the base of hate speech detection and algorithms allowed to be used. This would be based on the theory I propose that the free speech of an algorithm to condemn speech cannot override the free speech that the law protects if the Internet is a public space. This synthesis helps some problems in both fields. Computer science would give the tools to the law that would allow hate speech to found immediately as it is spoken and create clean environments that promote free speech. In the laws effort to promote free speech in the public forum, it would benefit from computer science to do so in the contemporary public forum, the internet. From the law discipline, there will be human moderators who won’t have the bias problem that was found in the computer science field. And there might be better moderation of memes that break the algorithms effectiveness such as distorted memes if the government is in on the sanctioning of online hate speech. The trick here would be how to give the government the information necessary to do this without compromising the private aspect of the internet users. Aside from this synthesis I see things either remaining the same where both fields serve as a check and balance to one another, or other extreme solutions that further isolate each discipline.

Bibliography

# Beauharnais v. Illinois, 343 U.S. 250 (1952)

# Brandenburg v. Ohio, 395 U.S. 444 (1969)

1. Elonis v. United States, 575 U.S. \_\_\_ (2015)
2. National Socialist Party of America v. Village of Skokie, [432 U.S. 43](https://supreme.justia.com/cases/federal/us/432/43/) (1977)

# R. A. V. v. St. Paul, 505 U.S. 377 (1992)

Badour, Jafar & Brown, Joseph, Hateful Memes Classification using Machine Learning, (2021)

1. Benjamin H. Winters, A Clear and Present Danger: The Need for Regulated Accountability for Online Service Providers to Preserve and Promote Free Speech, Notice and Due Process., (2018)

Das, Abhishek & Wahi, Japsimar & Li, Siyao, Detecting Hate Speech in Multi-modal Memes (2020)

Kiela, Douwe , “The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes, (2020)

1. Lily A. Coad, Compelling Code: A First Amendment Argument Against Requiring Political Neutrality in Online Moderation, (2021)
2. Mannirmal Kaur Jawa, The “Offensive” Oversimplification: An Argument for Hate Speech Laws in the Modern Era, (2021)
3. Michel Rosenfeld, Hate Speech in Constitutional Jurisprudence: A Comparative Analysis, 24 CARDOZO L. REV. 1523, (2003)

N.S Mullah and W.M.N.W. Zainon, Advances in Machine Learning Algorithms for Hate Speech Detection in Social Media: A Review, (2021)

Sabat, Benet Oriol, Hate Speech in Pixels: Detection of Offensive Memes towards Automatic Moderation, (2019)

Sap, Maarten, Card, Dallas, Gabriel, Saadia, Choi, Yejin, and Smith, A. Noah, The Risk of Racial Bias in Hate Speech Detection, (2019)

1. Velioglu, Riza & Rose, Jewgeni, Detecting Hate Speech in Memes Using Multimodal Deep Learning Approaches: Prize-winning solution to Hateful Memes Challenge (2020)
2. Warner, William & Hirschberg, Julia, Detecting hate speech on the world wide web (2012).
1. Michel Rosenfeld, Hate Speech in Constitutional Jurisprudence: A Comparative Analysis at 1530 [↑](#footnote-ref-1)
2. Id. at 1531 [↑](#footnote-ref-2)
3. Ibid. [↑](#footnote-ref-3)
4. Ibid. [↑](#footnote-ref-4)
5. Id. at 1532 [↑](#footnote-ref-5)
6. Michel, Rosenfeld at 1532 [↑](#footnote-ref-6)
7. Michel, Rosenfeld, at 1533 [↑](#footnote-ref-7)
8. Michel, Rosenfeld, at 1535 [↑](#footnote-ref-8)
9. Ibid. [↑](#footnote-ref-9)
10. Mannirmal Kaur Jawa, The “Offensive” Oversimplification: An Argument for Hate Speech Laws in the Modern Era, at 141 [↑](#footnote-ref-10)
11. Ibid. [↑](#footnote-ref-11)
12. *Beauharnais v. Illinois, 343 U.S. 250 (1952)* [↑](#footnote-ref-12)
13. Mannirmal Kaur Jawa at 142 [↑](#footnote-ref-13)
14. *Beauharnais v. Illinois, 343 U.S. 250 (1952)* [↑](#footnote-ref-14)
15. *Brandenburg v. Ohio, 395 U.S. 444 (1969)* [↑](#footnote-ref-15)
16. *Brandenburg v. Ohio, 395 U.S. 444 (1969)* [↑](#footnote-ref-16)
17. *National Socialist Party of America v. Village of Skokie,* [*432 U.S. 43*](https://supreme.justia.com/cases/federal/us/432/43/) *(1977)* [↑](#footnote-ref-17)
18. *National Socialist Party of America v. Village of Skokie,* [*432 U.S. 43*](https://supreme.justia.com/cases/federal/us/432/43/) *(1977)* [↑](#footnote-ref-18)
19. *R. A. V. v. St. Paul, 505 U.S. 377 (1992)* [↑](#footnote-ref-19)
20. *Elonis v. United States, 575 U.S. \_\_\_ (2015)* [↑](#footnote-ref-20)
21. Lily A. Coad, Compelling Code: A First Amendment Argument Against Requiring Political Neutrality in Online Moderation, at 469 [↑](#footnote-ref-21)
22. Id. at 457 [↑](#footnote-ref-22)
23. Benjamin H. Winters, A Clear and Present Danger: The Need for Regulated Accountability for Online Service Providers to Preserve and Promote Free Speech, Notice and Due Process, at 170 [↑](#footnote-ref-23)
24. Lily A. Coad, Compelling Code: A First Amendment Argument Against Requiring Political Neutrality in Online Moderation, at 485 [↑](#footnote-ref-24)
25. Velioglu, Riza & Rose, Jewgeni, Detecting Hate Speech in Memes Using Multimodal Deep Learning Approaches: Prize-winning solution to Hateful Memes Challenge, 1 (2020). [↑](#footnote-ref-25)
26. Warner, William & Hirschberg, Julia, Detecting hate speech on the world wide web, 20 (2012). [↑](#footnote-ref-26)
27. Ibid. [↑](#footnote-ref-27)
28. Ibid. [↑](#footnote-ref-28)
29. Ibid. [↑](#footnote-ref-29)
30. Badour, Jafar & Brown, Joseph, Hateful Memes Classification using Machine Learning, 4 (2021) [↑](#footnote-ref-30)
31. Warner, William & Hirschberg, Julia, Detecting hate speech on the world wide web, 20 (2012). [↑](#footnote-ref-31)
32. Ibid. [↑](#footnote-ref-32)
33. Ibid. [↑](#footnote-ref-33)
34. Warner, William & Hirschberg, Julia, Detecting hate speech on the world wide web, 20 (2012) [↑](#footnote-ref-34)
35. Warner, William & Hirschberg, Julia, Detecting hate speech on the world wide web 22, (2012) [↑](#footnote-ref-35)
36. Warner, William & Hirschberg, Julia, Detecting hate speech on the world wide web, 21 (2012) [↑](#footnote-ref-36)
37. N.S Mullah and W.M.N.W. Zainon, Advances in Machine Learning Algorithms for Hate Speech Detection in Social Media: A Review, 5 (2021) [↑](#footnote-ref-37)
38. Warner, William & Hirschberg, Julia, Detecting hate speech on the world wide web, 21 (2012) [↑](#footnote-ref-38)
39. Badour, Jafar & Brown, Joseph, Hateful Memes Classification using Machine Learning, 2 (2021) [↑](#footnote-ref-39)
40. Velioglu, Riza & Rose, Jewgeni, Detecting Hate Speech in Memes Using Multimodal Deep Learning Approaches: Prize-winning solution to Hateful Memes Challenge , 1 (2020) [↑](#footnote-ref-40)
41. Kiela, Douwe , “The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes, 3 (2020) [↑](#footnote-ref-41)
42. Ibid. [↑](#footnote-ref-42)
43. Sabat, Benet Oriol, Hate Speech in Pixels: Detection of Offensive Memes towards Automatic Moderation, 5 (2019) [↑](#footnote-ref-43)
44. Sabat, Benet Oriol, Hate Speech in Pixels: Detection of Offensive Memes towards Automatic Moderation , 4 (2019) [↑](#footnote-ref-44)
45. Ibid. [↑](#footnote-ref-45)
46. Sap, Maarten, Card, Dallas, Gabriel, Saadia, Choi, Yejin, and Smith, A. Noah, The Risk of Racial Bias in Hate Speech Detection, 2 (2019) [↑](#footnote-ref-46)
47. Ibid. [↑](#footnote-ref-47)
48. N.S Mullah and W.M.N.W. Zainon, at 88374 [↑](#footnote-ref-48)
49. Sap, Maarten, Card, Dallas, Gabriel, Saadia, Choi, Yejin, and Smith, A. Noah at 1 [↑](#footnote-ref-49)
50. Kiela, Douwe, at 8 [↑](#footnote-ref-50)
51. Badour, Jafar & Brown, Joseph, Hateful Memes Classification using Machine Learning , 2 (2021) [↑](#footnote-ref-51)
52. Sabat, Benet Oriol at 2 [↑](#footnote-ref-52)
53. Id. at 3 [↑](#footnote-ref-53)
54. Id.at 5 [↑](#footnote-ref-54)
55. Velioglu, Riza & Rose, Jewgeni. At 1 [↑](#footnote-ref-55)
56. Warner, William & Hirschberg, Julia, Detecting hate speech on the world wide web, 20 (2012). [↑](#footnote-ref-56)
57. Warner, William & Hirschberg, Julia, Detecting hate speech on the world wide web, 20 (2012) [↑](#footnote-ref-57)
58. Ibid. [↑](#footnote-ref-58)
59. Kiela, Douwe, “The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes, 3 (2020) [↑](#footnote-ref-59)