Exploring Disparities in Online Discourses:

The Impact of Search Filter Function on Outcomes across Multiple Platforms

Claire Jo, Hee Won Ahn, and Sehee Han

STA 220: Data & Web Technologies for Data Science

Dr. Peter Kramlinger

March 19, 2024

Author Note

Claire Jo Ph.D. Student in Communication

Hee Won Ahn M.S. Student in Statistics

Sehee Han M.S. Student in Statistics

Names are listed in alphabetical order according to the first name.

Abstract

This study investigates the impact of the search filter functions, specifically relevance and recency filters, on the representation of sensitive and socially problematic issues—suicide, abortion, and gun control—across three major digital platforms: YouTube, Reddit, and Yahoo. Employing a mixed-method approach, we scraped data related to these topics, applying Natural Language Processing (NLP) techniques such as TF-IDF, sentiment analysis, and topic modeling to analyze the collected datasets. Our findings reveal inconsistent differences in the portrayal of these issues across platforms, depending on the type of filter applied. While there was not a remarkable difference in result of difference sorting method for Youtube, there was a significant variation for Reddit. The outcomes of this research suggest the potential influence of the search filter function on shaping online discourses surrounding contentious social issues. By exploring the discrepancies introduced by different filtering mechanisms, this analysis suggests a broader conversation on algorithmic transparency and the need for more nuanced content moderation strategies that consider the complex dynamics of online communication and information consumption.

Exploring Disparities in Online Discourses:

The Impact of Search Filter Function on Outcomes across Multiple Platforms

1. Introduction

In the digital era, social media platforms such as YouTube, Reddit, and Yahoo have emerged as pivotal arenas for public discourse on sensitive issues and socially problematic topics. Specifically, these platforms significantly influence public opinion and societal norms through their search engine algorithms. However, the blind spots inherent in these algorithms can lead to a skewed representation of socially controversial issues such as suicide, abortion, and gun control. This research aims to uncover the extent to which the *relevance* and *recency (date)* search filters on these platforms affect the visibility and framing of these topics. By employing web scraping techniques to collect data across the three platforms, and applying Natural Language Processing (NLP) methods, including Term Frequency-Inverse Document Frequency (TF-IDF), sentiment analysis, and topic modeling, we provide a comprehensive analysis of the outcomes of each filter type. This analysis is important given the substantial social influence these platforms wield, potentially shaping public discourse and opinion in ways that may not fully represent the diversity of views on these contentious issues. Through this analysis, we expect to illuminate the implications of algorithmic filtering on public discourse, offering insights into the dynamics of digital platform moderation and its societal impacts.

As for the selection of keywords—suicide, abortion, and gun control—these topics hold significant relevance due to their societal impact on public health, individual freedoms, and community safety. In particular, these issues are highly contentious, sparking wide-ranging debates. Considering the polarizing natures of these topics, we hypothesized that the type of filter applied would yield significant disparities in search outcomes.

2. Data Acquisition

We collected data from YouTube, Reddit, and Yahoo News, focusing on search outcomes in mid-February 2024 with the following keywords: suicide, abortion, and gun control. For each platform, we employed two filters —recency and relevance— to collect data. We gathered 1,000 entries (comprising YouTube videos, Reddit posts, and Yahoo News articles) per filter, for each keyword. This resulted in a total dataset of 6,000 total entries, with an even split of 3,000 entries collected using the recency filter and 3,000 using the relevance filter. In addition to the initial dataset, our analysis also therefore retrieved comment data for each post across the platforms to capture audience responses to the selected posts. The volume of collected comments varied by platform, which is further explored in detail in the following section.

2.1. YouTube

Youtube is a platform that provides an online video sharing service which is one of the most popular platforms in the world, with more than 2.5 billions for people visiting everyday. Given its immense popularity and influence, our team decided to analyze people's reactions on keywords on Youtube. First, for YouTube video data collection, we utilized the YouTube Data API v3. With the API key, we created a web scraper that allows us to input search queries for which we want to gather metadata on search outcomes. We set the search filter to order results by "relevance" or "date". Then, we retrieved text-formatted metadata for each video, including video links, titles, descriptions, publication dates, durations, and channel names. Because the maximum result limit for each search page in our API queries was 50, we iterated over 20 pages to gather 1,000 entries for each keyword. However, the YouTube Data API v3 does not provide a direct means to collect video transcripts. Therefore,

we employed YouTubeTranscriptApi to gather the transcripts and then merged these with the previously collected metadata into a single dataset. Our next step was to collect user comments for each video in our dataset. Using the YouTube Data API v3 again, we made requests to YouTube's servers to fetch the first 20 comments from each video. We chose to limit the collection to 20 comments per video to manage the extensive volume of comments across 6,000 videos efficiently, preventing our dataset from becoming overwhelmingly large.

We have obtained:

(Abortion Relevance Text): 950 items, Number of Comments: 2043

(Abortion Date Text): 997 items, Number of Comments: 9385

(Gun Relevance Text): 800 items, Number of Comments: 1247

(Gun Date Text): 1000 items, Number of Comments: 22692

(Suicide Relevance Text): 1000 items, Number of Comments: 15733

(Suicide Date Text): 1000 items, Number of Comments: 15512

2.2. Reddit

Reddit is one of the social networks in America that's renowned for its rich and varied content encompassing links, text, images, video and more. It has been ranked 18th most visited websites in the world. Rich contents of Reddit include information on a broad spectrum of topics including social, gaming, politics and beyond. We anticipated that public opinion on certain social-related keywords (abortion, suicide, gun regulation) would be straightforward with honesty due to reddit’s characteristic of valuing internet privacy and anonymity compared to other platforms. People may not reveal their honest opinion on Youtube and Yahoo News due to its strict policy against opinion that is deemed inappropriate or offensive.Therefore, our team endeavored to obtain a dataset from Reddit.

Our team utilized the reddit API to obtain 1000 post text sets for each keyword and each sorting method, anticipating to obtain a total 6000 title, post text and corresponding

comments. However, due to Reddit API policy of limiting the number of dataset to extract, our team could not obtain 1000 each post text and comments. We have requested 20 times to Reddit , extracted 50 dataset in each request. However, compared to requesting 1000 posts at once, splitting requests resulted in a reduced number of contents that is obtained. This may be due to technical issues with the Reddit API, so we proceed our data acquisition by requesting 1000 dataset at once. We obtained data including Title, Post Text, ID, Score, Total Comments, Post URL, and Comments. Although our primary goal is to analyze the combination of Title and Post Test as a single unit and commit as another unit, we also obtained a score of contents to open possibilities of comparison with the result of our sentimental analysis. we have obtained:

(Abortion New Text): 249 items, Number of Comments: 1910 items
(Abortion Relevance Text): 232 items, Number of Comments: 27386 items
(Gun New Text): 441 items, Number of Comments: 3354 items
(Gun Relevance Text): 435 items, Number of Comments: 23177 items
(Suicide New Text): 494 items, Number of Comments: 1343 items
(Suicide Relevance Text): 452 items, Number of Comments: 40863 items

Our single unit of dataset includes at least over 200 content and does not have much variation, the number of comments in the relevant dataset surpasses over 20000, comments in dataset sorted by data only limit to 1000 to 3000 comments. This phenomenon could be expected since people tend to be exposed to keywords with relevance since it's default method of sorting contents of keywords.

2.3. Yahoo News

Yahoo is one of the most widely used search engines worldwide, providing various features beyond search, such as news, finance, email, and weather. In this analysis, Yahoo

News data was selected for the text analysis. While YouTube and Reddit attract a broader user base and contents about diverse topics, Yahoo News tends to have a relatively limited user base, with contents primarily focused on socioeconomic and political topics. We assumed that, even for the same topics, the terminology and expressions used on new websites would be more refined compared to the other two platforms. Therefore, Yahoo News data was chosen for conducting a comparative text analysis across platforms. Since Yahoo News selects articles from various news channels, we expect to observe diverse perspectives without leaning towards limited perspectives. Also, this platform was considered suitable for collecting all the news data at once because understanding the HTML syntax used by Yahoo News is sufficient for the analysis without having to individually understand the website structures of various news sources.

As Yahoo News does not provide an API, we collected data utilizing web scraping with Selenium. The scraping logic is as follows: first, we navigate to a specific category on the Yahoo News website and click on it. Then, we scroll until no more articles appear. After scrolling through all the pages, if an article's title or the beginning part of an article contains a specific keyword, its URL was collected. The URLs were opened one by one to scrape the titles and the text of the articles, storing them in a list.

Through preliminary research, we found that the articles relevant to some keywords appeared only in specific categories. Therefore, we collected articles from different categories for each keyword. For example, for "abortion" and "gun", the data was scraped from categories such as [us, politics, election], while for "suicide", the data was scraped from [us, health, the360] categories. Additionally, when searching for keywords on Yahoo News, using precise terms such as "gun control" or "gun regulation" yielded much fewer articles. Instead, searching for "gun" provided a broader range of articles, many of which were related to gun regulation due to the characteristics of the news data. We conducted scraping bi-daily over

the course of a week, collecting a total of 13 articles related to “abortion”, 19 articles related to “gun”, and 6 articles related to “suicide”, along with their headers and text. Despite the relatively small amount of text compared to other platforms, we found it sufficient to represent the characteristics of news data and therefore, proceeded with the analysis, despite not being able to collect comments data due to technical limitations.

we have obtained:

(Abortion Text): 13 items

(Gun Text): 19 items

(Suicide Text): 6 items,

3. Data processing

Our team conducted lower casing, tokenization, stopwords removal, steeming, and kept only letters written in Alphabet for all three media platforms for following reasons. Lower casing was conducted to ensure the consistency of data by treating each word with different cases as the same word. By Tokenization, we could extract features from the text data and keep a single word as the basic unit of analysis. We also undertook to remove stopwords that are shown often but do not have significant meaning (e.g. “is”, “are”, “the”). The process of stemming helped to normalize words by reducing them to their base form which led to enhanced consistency of the data. Lastly, Since our primary goal aligns with analyzing content written in English, we only kept letters in the Alphabet. Although most of the preprocessing methods were the same, we considered different stopwords removal for Youtube dataset and Reddit dataset. Since the Youtube dataset included a transcript of the video, it contained a significant amount of colloquial language. We included additional stopwords commonly found in transcripts such as "okay","yeah","well","uh”. For the reddit dataset, we ignored text and comments that include ‘I am a bot’, due to their automated

nature. Overall, by applying preprocessing techniques above, our team was able to clean and standardize the dataset, which made our dataset efficient and low-noise for applying NLP methods.

4. Analysis

Our team applied three NLP analysis methods, which are TFIDF, Sentimental Analysis, and Topic modeling to compare the results across different platforms and difference sorting methods (Relevance, Date).

We applied the TF-IDF method to represent the importance of words in the dataset and compare them with each platform. In the process of extracting frequency of words, we only obtained words with a sentimental score not equal to 0, to concentrate on key themes and sentiments without being distracted by neutral terms. With applying a sentimental score not equal to 0, our team was able to exclude common nouns that are not meaningful for our analysis.

For Sentimental Analysis, We applied VADER sentiment analysis tool, which assigns each word to positive, negative, and neutral sentient and combines scores to aggregate as an overall sentimental score of single text. The reason we applied this dictionary based model is not only because of its efficiency of handling large volumes of text data, but because of missing labels. Since our team did not have a label on our dataset, it was impossible for us to conduct a machine learning based model.

Lastly, Topic Modeling was conducted to identify specific subtopics within the documents searched by the same keywords. Topic Modeling is a technique that originated from the intuition that certain words would frequently appear in association with specific topics. We employed the LDA (Latent Dirichlet Allocation) to divide the documents into

subtopics for each keyword and determined the probability of each word being classified into each topic.

4.1. Interpretation of the Data

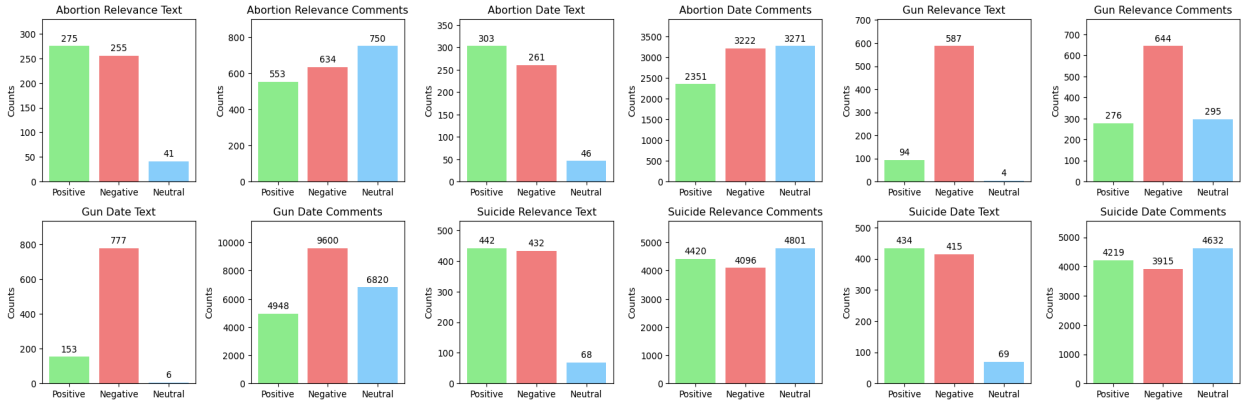
4.1.1. YouTube

TF-IDF

Text Title	Top 5 Words in Text	Comment Title	Top 5 Words in Comments
Abortion Relevance Text	[('care', 787), ('ban', 557), ('help', 516), ('support', 512), ('kill', 507)]	Abortion Relevance Comments	[('god', 180), ('kill', 127), ('murder', 125), ('love', 94), ('help', 73)]
Abortion Date Text	[('care', 695), ('ban', 563), ('support', 487), ('help', 471), ('god', 387)]	Abortion Date Comments	[('kill', 639), ('god', 589), ('murder', 496), ('care', 470), ('love', 323)]
Gun Relevance Text	[('gun', 12521), ('shoot', 2276), ('weapon', 1562), ('legal', 978), ('ban', 928)]	Gun Relevance Comments	[('gun', 806), ('protect', 85), ('kill', 85), ('shoot', 78), ('weapon', 77)]
Gun Date Text	[('gun', 17087), ('shoot', 2862), ('weapon', 2259), ('legal', 1378), ('ban', 1342)]	Gun Date Comments	[('gun', 8360), ('shoot', 1575), ('crime', 915), ('stop', 879), ('weapon', 846)]
Suicide Relevance Text	[('help', 1462), ('harm', 966), ('love', 944), ('kind', 920), ('god', 767)]	Suicide Relevance Comments	[('love', 1434), ('help', 1174), ('harm', 1006), ('cut', 770), ('stop', 538)]
Suicide Date Text	[('help', 1409), ('harm', 940), ('love', 913), ('kind', 878), ('god', 659)]	Suicide Date Comments	[('love', 1385), ('help', 1139), ('harm', 970), ('cut', 749), ('stop', 521)]

Interestingly, our analysis revealed no significant differences in search outcomes between using relevance or recency as the search filter. Even for the keyword "gun," the top-ranked results remained consistent across both relevance and recency filters. This may be due to certain characteristics of Youtube, since outcomes of Reddit vary with different search filters.

Sentiment Analysis



Youtube dataset shows no significant differences in outcomes of sentimental analysis which aligns with the result of TF-IDF.

In terms of Text dataset, results of sentimental analysis on text data reveal that there is not much difference between relevance and date in all three keywords. They both have similar proportions of positive, negative and neutral. Although, People's sentiment on Abortion and Sucidie is controversial, the ratio of positive and negative is approximately 50:50. However, we can conclude that people's opinion on Gun control is mostly negative.

Results of sentimental analysis on comment dataset does not show significant difference between sorting methods, However, the above graph shows that people's comment on abortion tends to show more negativity when sorted by date. However, the comments on abortion in Youtube are not as positive as Reddit. This may be due to the fact that content creators tend to focus more on controversial topics rather than personal experience, as seen on Reddit.

Topic Modeling

Text Title	Top Words in Text (Topic 1)	Top Words in Text (Topic 2)	Comment Title	Top Words in Comments (Topic 1)	Top Words in Comments (Topic 2)

Abortion Relevance Text	republican 0.997244 msnbc 0.996611 presid 0.995407 mississippi 0.994725 trump 0.994468 human 0.994418 overturn 0.994351 gun 0.993842 conserv 0.993694 governor 0.993668 Name: Topic 1, dtype: float64	ke 0.996632 mtp 0.993227 hai 0.993167 etv 0.993130 baad 0.992984 hindi 0.990537 cramp 0.989416 kit 0.989250 tablet 0.988508 kab 0.986245 Name: Topic 2, dtype: float64	Abortion Relevance Comments	life 0.997992 get 0.996974 god 0.996703 women 0.996641 dont 0.996270 child 0.996214 go 0.995947 think 0.995726 peopl 0.995693 would 0.995550 Name: Topic 1, dtype: float64	mam 0.993597 si 0.990816 hai 0.989921 ho 0.987446 la 0.987344 pl 0.986477 h 0.986342 kya 0.983646 se 0.983157 kit 0.982330 Name: Topic 2, dtype: float64
Abortion Date Text	elijah 0.990990 consciou 0.987398 comic 0.987260 jesu 0.986462 judi 0.986143 distinct 0.985136 bibl 0.984705 clown 0.984400 ben 0.984395 femin 0.982559 Name: Topic 1, dtype: float64	ke 0.996824 msnbc 0.996189 baad 0.995065 misoprostol 0.994407 nowthi 0.994162 hai 0.992705 etv 0.992687 abc 0.991063 nbc 0.990276 incomplet 0.990219 Name: Topic 2, dtype: float64	Abortion Date Comments	mam 0.998840 hai 0.998558 ho 0.998409 h 0.998090 kya 0.997655 ki 0.997413 ke 0.997326 se 0.997233 bleed 0.996988 din 0.996884 Name: Topic 1, dtype: float64	peopl 0.999527 women 0.999527 life 0.999524 dont 0.999434 get 0.999419 right 0.999413 child 0.999357 make 0.999219 woman 0.999172 say 0.999138 Name: Topic 2, dtype: float64
Gun Relevance Text	knife 0.996281 idaho 0.996207 knive 0.994631 nevada 0.994537 alaska 0.994272 cool 0.994081 reciproc 0.993937 stuff 0.993209 bro 0.993007 ffl 0.991856 Name: Topic 1, dtype: float64	nbc 0.996515 cb 0.991907 holyok 0.988576 snl 0.988482 lankford 0.986870 msnbc 0.986684 tatum 0.986418 msnbccom 0.978496 fnc 0.977899 nbcnewscom 0.977866 Name: Topic 2, dtype: float64	Gun Relevance Comments	peopl 0.995629 control 0.994875 protect 0.994616 gun 0.993270 think 0.991783 take 0.991652 crime 0.991170 owner 0.990650 crimin 0.990396 arm 0.990353 Name: Topic 1, dtype: float64	nevada 0.965546 reserv 0.947995 john 0.947577 mendicino 0.934507 washington 0.929552 rez 0.926819 wow 0.923179 conceal 0.922013 suit 0.921064 semi 0.920772 Name: Topic 2, dtype: float64
Gun Date Text	reciproc 0.994064 nonresid 0.991812 dakota 0.990018 bruin 0.988217 ccn 0.987873 injunct 0.987460 ccw 0.987307 nevada 0.985213 roadsid 0.984625 wildlif 0.983752 Name: Topic 1, dtype: float64	kid 0.999195 australia 0.998339 kill 0.998126 lobbi 0.997603 oh 0.997421 liter 0.997385 god 0.997346 murder 0.997265 democrat 0.997228 black 0.996997 Name: Topic 2, dtype: float64	Gun Date Comments	gun 0.999917 peopl 0.999790 law 0.999774 dont 0.999676 make 0.999477 firearm 0.999441 would 0.999415 carri 0.999334 crime 0.999268 think 0.999266 Name: Topic 1, dtype: float64	ha 0.998824 de 0.996603 jesu 0.996120 que 0.995175 lo 0.994884 e 0.994278 bless 0.993401 prayer 0.993368 trump 0.992562 v 0.991857 Name: Topic 2, dtype: float64
Suicide Relevance Text	realli 0.999334 um 0.999235 someth 0.998915 kind 0.998736 thing 0.998718 ive 0.998655 god 0.998637 know 0.998457 emot 0.998440 way 0.998386 Name: Topic 1, dtype: float64	yang 0.992507 di 0.989769 ini 0.988835 diri 0.988059 untuk 0.987189 sendiri 0.986443 dari 0.983983 dengan 0.982775 perilaku 0.980314 autogen 0.980028 Name: Topic 2, dtype: float64	Suicide Relevance Comments	de 0.999162 que 0.998903 la 0.998815 le 0.998635 je 0.997758 en 0.997577 et 0.997546 mi 0.997376 se 0.997110 pa 0.997108 Name: Topic 1, dtype: float64	im 0.999809 feel 0.999696 dont 0.999632 know 0.999622 get 0.999616 help 0.999599 one 0.999591 peopl 0.999568 realli 0.999553 make 0.999522 Name: Topic 2, dtype: float64

Suicide Date Text	cid 0.993177 lantern 0.991454 robbi 0.991394 margot 0.991350 gunn 0.990553 il 0.988932 ayer 0.988103 di 0.987963 un 0.987828 galaxi 0.987791 Name: Topic 1, dtype: float64	god 0.998728 word 0.998518 understand 0.998369 scar 0.998227 said 0.998165 realli 0.998131 ive 0.998105 physic 0.997926 someon 0.997879 hurt 0.997856 Name: Topic 2, dtype: float64	Suicide Date Comments	de 0.999513 le 0.999276 la 0.999193 que 0.999167 et 0.998726 je 0.998549 en 0.998430 pa 0.998345 un 0.998053 est 0.998020 Name: Topic 1, dtype: float64	feel 0.999610 know 0.999540 get 0.999537 peopl 0.999496 im 0.999495 one 0.999490 help 0.999466 love 0.999446 make 0.999436 dont 0.999423 Name: Topic 2, dtype: float64

When topic modeling was applied to YouTube data, similar to the previous analysis, meaningful modeling was not achieved. The main reason is that YouTube is a worldwide platform, leading to many comments being written in languages other than English. Additionally, the text in the main body is often automatically generated from human voice to script, resulting in the inclusion of onomatopoeia, exclamations, redundant expressions, or non-English expressions in text analysis. It is expected that meaningful results will be obtained in future analyses if there is an algorithm to filter out non-English expressions represented in alphabets.

4.1.2. Reddit

TF-IDF

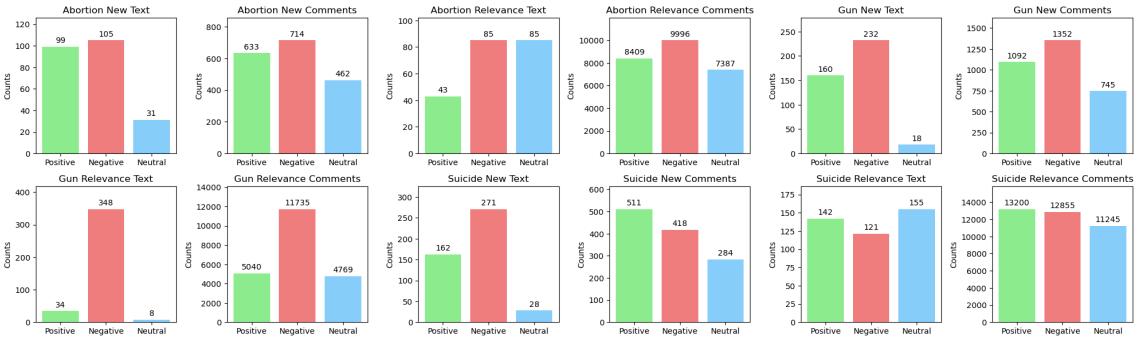
Text Title	Top 5 Words in Text	Comment Title	Top 5 Words in Comments
Abortion New Text	[('help', 66), ('support', 61), ('love', 56), ('friend', 55), ('well', 54)]	Abortion New Comments	[('care', 105), ('help', 93), ('well', 91), ('sure', 80), ('best', 80)]
Abortion Relevance Text	[('ban', 36), ('help', 29), ('friend', 28), ('support', 23), ('protect', 18)]	Abortion Relevance Comments	[('care', 1388), ('support', 1072), ('love', 1058), ('well', 983), ('god', 932)]
Gun New Text	[('gun', 834), ('weapon', 513), ('kill', 405), ('well', 376), ('fire', 368)]	Gun New Comments	[('gun', 1363), ('shoot', 281), ('well', 269), ('problem', 177), ('better', 164)]
Gun Relevance Text	[('gun', 798), ('shoot', 82), ('weapon', 51), ('ban', 40), ('argument', 29)]	Gun Relevance Comments	[('gun', 16125), ('shoot', 2419), ('ban', 1629), ('problem', 1440), ('weapon', 1408)]
Suicide New Text	[('help', 293), ('friend', 285), ('harm', 258), ('love', 217), ('bad', 187)]	Suicide New Comments	[('help', 161), ('better', 104), ('bad', 87), ('hope', 79), ('love', 76)]
Suicide Relevance Text	[('cut', 55), ('harm', 54), ('help', 52), ('friend', 45), ('kill', 41)]	Suicide Relevance Comments	[('help', 2493), ('love', 1847), ('harm', 1681), ('better', 1599), ('well', 1472)]

Interestingly, compared to youtube, our analysis revealed significant differences in search outcomes between using relevance or recency as the search filter.

In terms of text dataset, abortion sorted by date tend to show more positive and encouraging words compared to sorted relevance. Words “help” were shown the most when sorted by date, which suggest that posts are requesting assistance or guidance. On the other hand, “ban” was shown the most when sorted by relevance, indicating that people focus on abortion-related-policies. Keyword Gun does not show much difference, However, when sorted by relevance “ban”,”argument” was one of the words that's commonly used, indicating that people focus on gun control policies. These indicate that sorting relevance results in more policies related posts.

Comment dataset revealed the same phenomenon compared to text dataset, when sorted by relevance “ban” was also one of the words that's commonly used, indicating people’s interest in gun control policies.

Sentiment Analysis



The Reddit dataset shows significant differences in outcomes of sentimental analysis compared to Youtube dataset. One of the key differences is how the text dataset and comment dataset have different sentimental. In terms of sorted Suicide and Abortion text by recent and

relevance, there appears to be mostly negative sentiments in main posts related to suicide and abortion, while comments shows more positive sentiments. This suggests that many individuals posting are experiencing hardship, and comments are offering them support for abortion and suicide.

Both YouTube and Reddit exhibit similar distributions of sentiment between the main posts and comments of gun control. Main posts tend to be mostly negative, while comments show slightly more positivity. Further investigation is needed to understand the underlying reasons for this pattern. Our assumption is that people tend to debate on gun control with a minority of people exhibiting positive sentiment on gun control.

Topic Modeling

Text Title	Top Words in Text (Topic 1)	Top Words in Text (Topic 2)	Comment Title	Top Words in Comments (Topic 1)	Top Words in Comments (Topic 2)
Abortion New Text	sourc 0.978500 republican 0.976544 yakuza 0.974895 trump 0.974082 magic 0.969866 vote 0.968542 calli 0.968016 eli 0.967816 biden 0.967789 witch 0.966144 Name: Topic 1, dtype: float64	advic 0.981720 partner 0.980718 fianc 0.979882 bc 0.979601 pill 0.978838 period 0.975653 th 0.974956 test 0.974556 doctor 0.972493 se 0.971922 Name: Topic 2, dtype: float64	Abortion New Comments	gop 0.987962 ban 0.983633 govern 0.983473 ivf 0.978236 biden 0.973348 poor 0.973196 trump 0.972955 senat 0.970873 republican 0.967962 gun 0.967045 Name: Topic 1, dtype: float64	feel 0.994168 relationship 0.991001 end 0.989781 week 0.989552 wouldnt 0.989439 sorri 0.989170 best 0.986709 date 0.985864 pain 0.985311 option 0.984735 Name: Topic 2, dtype: float64
Abortion Relevance Text	antiabort 0.926724 protest 0.909637 approv 0.848848 alabama 0.841647 jd 0.840899 nationwid 0.839889 fetu 0.834700 walker 0.834260 voter 0.832259 enshrin 0.817491 Name: Topic 1, dtype: float64	get 0.988123 know 0.986570 time 0.984972 dont 0.984775 husband 0.984527 told 0.983324 think 0.982685 say 0.982092 mother 0.982088 day 0.981787 Name: Topic 2, dtype: float64	Abortion Relevance Comments	love 0.999092 husband 0.998773 na 0.998608 friend 0.998431 sorri 0.998331 plea 0.998014 op 0.997826 divorc 0.997763 proud 0.997744 glad 0.997669 Name: Topic 1, dtype: float64	state 0.999735 republican 0.999656 vote 0.999578 law 0.999528 ban 0.999336 birth 0.999326 govern 0.999129 conserv 0.999127 legal 0.999092 fetu 0.999016 Name: Topic 2, dtype: float64

Gun New Text	abort 0.996950 firearm 0.996003 dmg 0.995994 legal 0.995672 suprem 0.995623 accuraci 0.995459 constitut 0.995378 howev 0.995145 scam 0.995138 terran 0.994632 Name: Topic 1, dtype: float64	cib 0.993997 lp 0.993409 remix 0.990661 gen 0.990594 pepper 0.990169 tenaci 0.990028 volbeat 0.989928 dogg 0.989819 mario 0.989447 vgvg 0.988487 Name: Topic 2, dtype: float64	Gun New Comments	gun 0.999395 state 0.998593 law 0.998556 peopl 0.997690 firearm 0.997521 right 0.997472 shoot 0.997274 school 0.996868 militia 0.996669 regul 0.996393 Name: Topic 1, dtype: float64	chapter 0.994378 artifici 0.989489 hivemind 0.989296 waffl 0.981386 song 0.981200 dog 0.980299 pokemon 0.979059 combat 0.978502 fan 0.977550 roblox 0.975197 Name: Topic 2, dtype: float64
Gun Relevance Text	shoot 0.993866 regul 0.992724 violenc 0.991093 ghost 0.989349 weapon 0.989055 mass 0.987981 peopl 0.987724 dont 0.987110 would 0.986999 done 0.986241 Name: Topic 1, dtype: float64	judg 0.913048 declin 0.887559 congress 0.866455 dealer 0.862033 trump 0.851661 washington 0.849559 protest 0.844788 uk 0.842652 quadrant 0.835463 guncontrol 0.831056 Name: Topic 2, dtype: float64	Gun Relevance Comments	firearm 0.999700 law 0.999691 violenc 0.999484 crime 0.999446 crimin 0.999321 alcohol 0.999315 health 0.999295 check 0.999293 mental 0.999209 illeg 0.999189 Name: Topic 1, dtype: float64	fuck 0.999401 republican 0.999106 vote 0.998998 democrat 0.998411 shit 0.998357 trump 0.997850 elect 0.997762 conserv 0.997673 parti 0.997613 teacher 0.997428 Name: Topic 2, dtype: float64
Suicide New Text	hd 0.994038 batteri 0.982927 shane 0.982678 itun 0.982312 phone 0.981977 therapist 0.980901 rick 0.980752 gp 0.979660 uv 0.978737 idk 0.977670 Name: Topic 1, dtype: float64	wisdom 0.990927 suzu 0.989154 emblem 0.989000 okab 0.984024 crime 0.983016 muse 0.982270 tome 0.981298 weapon 0.980482 user 0.980168 mass 0.979819 Name: Topic 2, dtype: float64	Suicide New Comments	pageant 0.969065 healer 0.968674 john 0.965730 ban 0.961224 shuffl 0.956480 kashimo 0.950073 gue 0.948090 lfg 0.946677 soldier 0.945955 lobbi 0.945562 Name: Topic 1, dtype: float64	help 0.996570 feel 0.995925 thing 0.995413 take 0.993982 give 0.992775 sorri 0.992281 go 0.992004 live 0.991198 know 0.991178 sound 0.991016 Name: Topic 2, dtype: float64
Suicide Relevance Text	xb 0.988092 squad 0.975914 rate 0.974540 smell 0.969877 justic 0.968484 leagu 0.967949 player 0.963456 collect 0.960700 rank 0.956619 rocksteadi 0.955293 Name: Topic 1, dtype: float64	scar 0.985201 wife 0.972068 got 0.971054 selfharm 0.970573 cut 0.967451 post 0.966582 ill 0.965099 urg 0.963430 arm 0.963228 alway 0.961417 Name: Topic 2, dtype: float64	Suicide Relevance Comments	help 0.999760 feel 0.999746 someon 0.999662 life 0.999651 self 0.999527 harm 0.999522 friend 0.999522 depress 0.999466 mental 0.999466 talk 0.999457 Name: Topic 1, dtype: float64	game 0.999713 movi 0.999410 squad 0.999268 rate 0.999042 countri 0.998688 charact 0.998583 arkham 0.998303 harley 0.998282 dc 0.998143 batman 0.998138 Name: Topic 2, dtype: float64

When examining the results of topic modeling on Reddit data with the keyword "*abortion*," it can be observed that it is divided into two main topics of political topic and lifestyle topic. In the political topic, the words that are most likely to appear are "*republican, trump, vote, biden, govern*", and there is not much difference between relevance and recency filters. However, there is a difference between the two filters for the lifestyle topic. Under the recency filter, frequent words in posts include "*partner, fiancé, pill, period, doctor*", while comments include "*feel, relationship, wouldn't, pain*". Under the recency filter, posts are

mainly from people who have become pregnant without preparation and have had an abortion, and comments mainly provide words of comfort. In contrast, under the relevance filter, words that are most likely to appear include "*husband, mother, time, think*", and comments also include "*love, husband, friend, divorce*". Comparing the two filters, for the keyword "abortion," it can be seen that more provocative posts written in unstable situations are found under the recency filter. For the keyword "gun," under the relevance filter, it is divided into a political topic ("*legal, supreme, constitute, state*") and a topic related to shooting incidents ("*shoot, weapon, alcohol, mental*"), while under the recency filter, it is divided into a political topic and a miscellaneous topic ("*dog, volbeat, mario, pokemon, fan, roblox*") that are unrelated to gun issues . For the keyword "suicide," under the relevance filter, it is clearly divided into a topic related to real suicide concerns ("*scar, self-harm, cut*") and a topic related to games ("*squad, league, player*"), but under the recency filter, words related to self-harm appear in both topics, and there are many seemingly unrelated words, making it difficult to make a clear distinction between two topics.

4.1.3. Yahoo

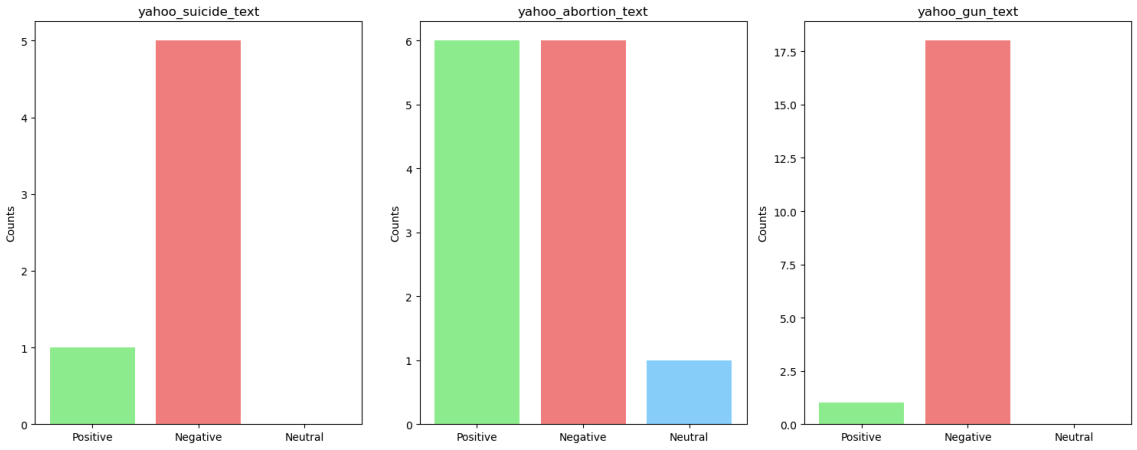
TF-IDF

Text Title	Top 5 Words in Text
yahoo_suicide_text	[('care', 9), ('help', 8), ('kill', 8), ('gun', 8), ('allow', 7)]
yahoo_abortion_text	[('support', 55), ('ban', 50), ('protect', 38), ('care', 28), ('legal', 18)]
yahoo_gun_text	[('gun', 75), ('fire', 19), ('weapon', 18), ('ban', 17), ('victim', 12)]

Although we could not obtain a large volume of datasat to fully explore characteristics of Yahoo news text, Yahoo news appears to have numerous articles related to policy. Since it's a news platform rather than social media, it tends to cover political matters more extensively.

Higher proportion of police-related words such as “allow” ,“legal” and ”ban” appeared more frequently than social media when extracting articles related to gun control and abortion. The term ‘support’ appeared as a result of TF-IDF across all three platforms, indicating the social atmosphere of supporting abortion regardless of social media and news platforms.

Sentiment Analysis



Even though we didn’t obtain numerous dataset from Yahoo news, abortion and gun dataset shows a similar distribution of positive and negative sentiment compared to Youtube and Reddit, where abortion has a 1:1 ratio, and mainly negative sentiment on gun control. This shows that all three platforms show controversial sentiment on keyword ‘abortion’, on the other hand, mainly negative sentiment on keyword ‘gun control’ which can be used to investigate public opinion .

Topic Modeling

Text Title	Top Words in Text (Topic 1)	Top Words in Text (Topic 2)

yahoo_suicide_text	media 0.968106 social 0.964659 mental 0.943300 would 0.940413 children 0.930339 gooden 0.929852 bill 0.925415 say 0.919582 law 0.915473 torr 0.914259 Name: Topic 1, dtype: float64	nex 0.857451 benedict 0.836439 fight 0.830775 publish 0.822245 macron 0.821140 relea 0.820375 administ 0.812004 depart 0.788024 origin 0.782814 owasso 0.782059 Name: Topic 2, dtype: float64
yahoo_abortion_text	ivf 0.953793 genderaffirm 0.928683 alabama 0.910842 stabil 0.900591 taylor 0.899123 embryo 0.895608 sonza 0.889187 introduc 0.886160 fair 0.884473 chief 0.881237 Name: Topic 1, dtype: float64	burkhart 0.966247 tiller 0.947111 wyom 0.946055 staff 0.923877 casper 0.919412 minnesota 0.918419 home 0.917114 wichita 0.916702 video 0.915578 clinic 0.908990 Name: Topic 2, dtype: float64
yahoo_gun_text	son 0.943350 trial 0.924031 man 0.920680 mobil 0.918283 king 0.916942 west 0.912158 california 0.905566 distefano 0.903905 deal 0.901939 hollywood 0.900856 Name: Topic 1, dtype: float64	bump 0.978240 stock 0.976589 banuelo 0.962568 ban 0.959934 bill 0.943955 suprem 0.936543 atf 0.934493 rifl 0.923466 trigger 0.919602 event 0.916181 Name: Topic 2, dtype: float64

Since most articles on Yahoo News are related to incidents, politics, and socio-economic topics, they appear to contain similar contents. However, through topic modeling, it was possible to identify recent emerging subtopics within main topics filtered by the same keyword. For the keyword "abortion," topics were identified concerning states that recently banned abortion ("*Wyoming, Minnesota*") and issues related to current themes of discourse surrounding abortion ("*IVF, gender affirmation, embryo*"). For the keyword "gun," the subtopics of articles were sharply divided. Topic 1 focused on gun accidents ("*trial, man, Hollywood, deal*"), while topic 2 had a higher probability of words related to gun bump stocks ("*bump, stock, supreme*"). Similarly, for the keyword "suicide," clear topics were identified. Topic 1 focused on the impact of social media on mental health of teenagers ("*mental, social, media, children*"), while topic 2, indicated by political terms ("*Macron,*

administration, department"), suggested that a politician may have recently committed suicide or proposed political agendas related to suicide. Yahoo News articles are characterized by refined terminology and the absence of slang or meme-related words in the text, making it relatively easier to divide and compare topics.

4.2. Challenges

One significant challenge in our research was the need to collect data from three distinct platforms, each with its own unique structure. This required the development of multiple web scrapers, each tailored to the specific format of the platform it targeted. The complexity of adapting to these varying structures was substantial, yet it offered us a valuable learning experience in data collection across diverse web environments.

Second, the nature of discourse varied significantly across the platforms, adding another layer of complexity to our analysis. For example, the language used in YouTube video metadata and comments tended to be more colloquial, reflecting a wide range of "less filtered" audience expressions. Conversely, Reddit posts and comments were generally focus on personal experiences compared to YouTube, known for its personal and candid content. Regarding Yahoo News, given the structured format of journalistic articles, the featured language was highly refined and formal. In this regard, we had to employ different data preprocessing techniques for each platform and develop a sophisticated approach for removing stopwords.

Third, due to the dynamic nature of comments data within Yahoo News's website frame, we encountered difficulties in collecting data using conventional HTML syntax and Selenium. Considering the technical limitations, it was deemed challenging to collect comments data, so only the main article text from Yahoo News was utilized. Therefore, the

focus was placed on analyzing the main text rather than examining the differences between relevance and recency filters of comments in Yahoo News. Our significance was placed on revealing differences between Yahoo News and other platforms through text analysis.

5. Conclusion

To sum up, we conducted text analysis by collecting posts and comments related to provocative keywords from three platforms: YouTube, Reddit, and Yahoo News. We applied three techniques—TF-IDF, sentiment analysis, and topic modeling—then analyzed: 1) differences across platforms and 2) differences between texts when applying relevance and recency filters. Across all platforms, a common characteristic was that applying the relevance filter resulted in more political and social posts and comments compared to the recency filter. Notably, when conducting sentiment analysis, there was a clear difference between the recency and relevance filters for Reddit. Reddit comments tend to be more positive compared to YouTube, so if a user seeks validation and support for a certain experience and situation, Reddit might be the better platform to utilize. Additionally, sentiment varied for each keyword across platforms, enabling analysis of the predominant emotions users have towards specific topics on each platform.

We anticipate expanding the scope of data or conducting more rigorous text analysis in future analyses. Through platform-specific and filter-specific text analyses, we can suggest to users which platform to use for specific purposes and improve platform recency and relevance algorithms based on these results. Furthermore, we anticipate utilizing this analysis for marketing or political purposes in the future.