

# Breaking through Inequality of Information Acquisition among Social Classes: A Modest Effort on Measuring "Fun"

Chenghao Xiao<sup>★</sup> Baicheng Sun<sup>♠</sup> Jindi Wang<sup>★</sup> Mingyue Liu<sup>★</sup> Jiayi Feng<sup>♦</sup>

<sup>★</sup> Department of Computer Science, Durham University

<sup>♠</sup> School of Social Science, Tsinghua University

<sup>♦</sup> Beijing Jiaotong University

[chenghao.xiao@durham.ac.uk](mailto:chenghao.xiao@durham.ac.uk)

## Abstract

With the identification of the inequality encoded in information acquisition among social classes, we propose to leverage a powerful concept that has never been studied as a linguistic construct, "fun", to deconstruct the inequality. Inspired by theories in sociology, we draw connection between social class and information cocoon, through the lens of fun, and hypothesize the measurement of "how fun one's dominating social cocoon is" to be an indicator of the social class of an individual. Following this, we propose an NLP framework to combat the issue by measuring how fun one's information cocoon is, and empower individuals to emancipate from their trapped cocoons. We position our work to be a domain-agnostic framework that can be deployed in a lot of downstream cases, and is one that aims to deconstruct, as opposed to reinforcing, the traditional social structure of beneficiaries (Jin et al., 2021).

## 1 Introduction

*Does a researcher necessarily want to be surrounded by research-related content at any time of a day?*

*Would under-privileged members in society be aware if they are consuming entertainment content all the time?*

This paper starts with posing the above questions on two extreme cases, which indicate a misalignment of a (content, concept) pair that members of different social classes are stuck in, during the process of information acquisition.

While under-privileged social class is identified to be trapped in entertainment content (Xu et al., 2020), which in turn reinforces their social class; higher social class is prone to content that causes anxiety, especially during the periods of a day that members in this class are in urgent need of escaping from their social roles and mental burdens (Wang, 1999; Oh and Pham, 2022), while trapped in highly

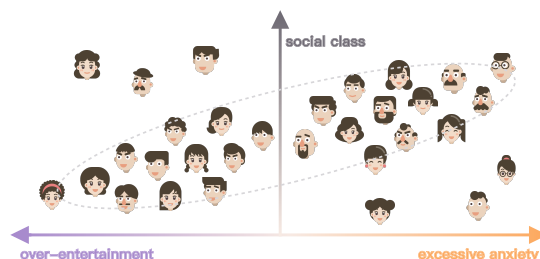


Figure 1: Theoretical Relation between Fun and Social Class Status Depicted in Social Science (Oh and Pham, 2022; Xu et al., 2020; Wang, 1999)

anxious content due to the preference they have shown to the algorithms, attributable to their social roles.

We argue that there exists inequality of information acquisition among social classes, through which a powerful concept we leverage might be able to help interpret: "fun" – a construct that has never been studied by the NLP community, or never studied as a *linguistic* construct at all, and one that we try to distinguish from "humor", with the latter having been heavily investigated in NLP research.

In this work, we draw upon deconstructing the inequality among social classes, in combat with the issue that advancements of NLP techniques are in fact sometimes reinforcing traditional social structure of beneficiaries (Jin et al., 2021) in society - an evaluation heuristic proposed recently by Jin et al. (2021) on aligning NLP with social good.

Moreover, we aim not to only address the inequality posed with social groups that are under-privileged socio-economically. We in turn believe that every social class is under-privileged in different ways, with an example discussed above regarding some social groups that are typically deemed privileged oftentimes do not actually have the privilege to liberate from their social roles (Oh and Pham, 2022). We propose, by understanding a key concept in interpreting social cocoons - *how fun a social cocoon is* - we could facilitate and empower

members of different social classes to emancipate from their cocoons based on their needs.

The paper is structured as follows: We first turn to theories in sociology (Section 2) to draw connection between social class and information cocoon, through the lens of "fun" - an under-studied concept, and introduce our NLP framework as an external regularization to confront the issues involved. Section 3 introduces the first part of our framework by training an intermediate language model to understand the concept of fun at a population level. Following that, Section 4 shows the strong utility of the intermediate model in terms of its transferability in downstream tasks, corroborated respectively in zero-shot and cross-domain user-aware fine-tuning setups. Moreover, in Section 4 we not only propose some real-life cases to deploy our framework, but also conduct user studies to align their perception with our intermediate model. We consider our work to be a framework that mitigates the Matthew effect of inequality in information acquisition in traditional social structure (Jin et al., 2021) and put "NLP for social good" into practice.

## 2 "Fun" Framework

In this section, we sketch the thinking heuristics of our framework by basing it on the grounds of sociology theories, drawing connection between social class and information cocoon, through the lens of "fun".

We theoretically reason that, the measurement of "how fun a social cocoon is", has been a strong indicator of the social class, or the under-privileged position, that an individual is stuck in, and thus is a metric that shows strong utility toward helping individuals of different social classes to combat the ubiquitous inequality in information acquisition.

Lastly, we introduce the "Fun" framework, enabling members of a specific social group to escape to the other side of this metric with an NLP model.

### 2.1 Concepts

**Information cocoon** Do we feel at ease after sporadically spending a whole day online? A positive answer may not be easily drawn, for the existence of algorithms and that big data is increasingly turning itself from servicing to controlling - a process in which stands information cocoon as a major consequence.

People are gradually losing their subjectivity and

degrading to complete recipients of information. The perniciousness comes right from this unconscious process of being besieged by a closed information system, making individuals more vulnerable and less adaptable to unfamiliar topics or conflicting views (Sunstein, 2007). Thus, it is likely to incur extremism and political polarization in terms of mass-level behavior (Barberá et al., 2015), which is often provocative and incivility-prone (Gervais, 2015).

Under micro-narratives, information cocoon pertains to one's everyday experience and actual social network. Sociological studies have shown that, even before the invention of the Internet, humans had already been unconsciously restricted by materiality and social relationships in the real world (Fei et al., 1992). People are largely influenced by their primary groups, i.e., kinships, close friends, and neighbors (Cooley, 1955). In extreme circumstances, primary groups set the boundary that one can reach to in their lifetime. Considering the empirical evidence that the stronger the tie connecting two individuals, the more similar they will be (Granovetter, 1973), primary groups have in fact formed a pre-Internet information cocoon, where individuals develop their mindsets and habits in similar ways. However, instead of strong ties, weak ties are more useful in expanding the flow of information (Granovetter, 1973, 2018, e.g., job-seeking, political mobility). This further indicates the significance of diversity and expansion of relationships when transferring and acquiring information.

Thus, it is only fair to say that the Internet constructs an unprecedentedly personalized yet information-intense space through algorithms, making information cocoon a non-negligible problem. In the past, it was relatively safe for people to be immersed in similar ideas, as they were not likely to build connections beyond their primary groups and accordingly have less risks confronting opposition and information overload. By contrast, the combination of the Internet and recommender systems nowadays forges a dilemma: users are, on the one hand, empowered to jump out of their primary groups, yet on the other hand, **thrown into a digital primary group**, in which information is passively sent, not actively sought.

Retrieved to the earliest definition by Sunstein (2006), the negative effect of information cocoon is mainly the illusionary friendliness and comfort brought by like-minded opinions and homogeneous

messages (Sunstein, 2006). We must concede that digital information acquisition nowadays fails to keep its technological promises (Harambam et al., 2018), for accurately identifying users' needs and offering customized information. In other words, a familiar environment, created by the "sorting" mechanism, cannot guarantee the bottom line of being harmless to the Internet users (Gervais, 2015).

**Fun: A taken-for-granted yet understudied concept** To counterpoise the negative feelings caused by information cocoon and empower individuals to retrieve their subjectivity, we turn to "fun", a taken-for-granted yet understudied concept, for help, especially during a time when we have seen the cognitive reduction of fun led by COVID-19 (Oh and Pham, 2022) and the deterioration of on-line environment<sup>1</sup>. Fun pertains to almost every aspect of our life, yet never gets serious attention. It soothes everyday tension and seasons bland normalcy. However, even among the very little research that mentioned fun as an independent concept, namely psychological investigation of fun (Oh and Pham, 2022) and physical education studies (MacPhail et al., 2008; Bengoechea et al., 2004; Scanlan and Simons, 1992), fun is related to distraction, reduced to the by-product of other social facts, or used interchangeably with happiness, well-being, enjoyment, sense of achievement, deviance, and humor (Fincham, 2016). The sophisticated crossover amidst these positive affective states makes it hard to theorize fun as an independent phenomenon (Fincham, 2016; Blythe et al., 2004).

Besides, though we all crave for fun, it is not seen as an essential factor as water is to humans. We tend not to exchange things deemed as more secularly important for fun. For instance, few parents would allow their kids to stay at home and have unbridled fun simply because fun is marginalized by regimentations in schools. Besides, fun is too contingent to handle, for having too much fun is sometimes frivolous and frowned upon (Fincham, 2016; Goffman, 1961). Additionally, to theorize fun sounds contradictory to its quotidian nature. Consequently, there is no specific and distinctive branch in sociology that sets its goal to understand fun theoretically until the advent of *The Sociology of Fun* by Ben Fincham in 2016 (Fincham, 2016).

It is worth mentioning that fun is a broader concept compared to humor, in spite of their sim-

ilarities (Ruch, 2001; Fincham, 2016). Fun is highly contextual and diverse, while humor is simply more about making people laugh (Martin and Ford, 2018). Thus, jokes are important sources for studies aiming to detect humor (Yang et al., 2021; Weller and Seppi, 2019). By contrast, laughter is not necessary to make people feel fun (Fincham, 2016). From a sociological perspective, humor is more hierarchical, while fun requires equality to take place (Podilchak, 1991). The semantic archaeology of fun has shown its intertwining history with **social class, judgement and transgression** (Blythe and Hassenzahl, 2018), revealing its rebellious nature and tendency toward equality. This not only distinguishes itself from other emotion-related concepts, but also add up to the legitimacy of using "fun" as a weapon to combat the inequality caused by information cocoons. Fun is not that simple but digs deeper into human's mental need for participation and freedom. It is common for both a person who has just finished a book and a couple playing badminton to feel fun (Oh and Pham, 2022).

In a word, what really matters is not trying to universalize people's experience of fun, but to uncover the common mechanism of cultivating fun. In order not to be confused by the countless realizations of fun in real life, psychological studies become inspiring, for they try to offer the fundamental pillars of the mechanism to produce fun, among which stands out the definition by Oh and Pham (2022): **an experience of liberating engagement**.

## 2.2 Tackling Inequality of Information Acquisition through the Lens of "Fun"

To this point, it might have already been clear that we are to leverage the concept of fun as a conceptual weapon to provide individuals with freedom to move along their information cocoons, which to different extent, indicates the social classes, or under-privileged positions that they are in.

By its very nature of geographic space, one can easily associate information cocoon with semantic space from the perspective of natural language processing. However, it is identified that existing endeavors that utilize language models, especially the state-of-the-art contextualized ones such as BERT (Devlin et al., 2019) and their variants like Sentence Transformers (Reimers and Gurevych, 2019) to study information cocoon, mostly concern political polarization (Jiang et al., 2021), news and items recommendation (Shi et al., 2021; Song et al.,

<sup>1</sup><https://www.microsoft.com/en-us/online-safety/digital-civility>

2022), and the spread of Covid-19 misinformation (Röchert et al., 2021), while few studied information cocoons in cultural consumption (Xu et al., 2020). Xu et al. (2020) leveraged word embedding models to analyze information cocoon in digital media, with the purpose of studying information cocoon as a cultural space and its relationship with social class, indicating that the disadvantages of vulnerable groups in the process of acquiring knowledge may further widen social inequality.

In the quest for information equality among social classes, we propose that formulating the implicit "geographical space" expressed in the information cocoon as a continuous representation of "fun" is not only capable of addressing the **anxiety issues posed with higher social class** but is also a panacea for helping **under-privileged social class escape from their "knowledge-absent" cocoons**, by adjusting their scale of "fun" to attain/filter knowledgeable recommendations, instead of being stuck in entertainment content (Xu et al., 2020).

We introduce a computational framework for measuring fun in language. This framework can be understood in a two-step setup: a) intermediate pre-training at a population level. b) Individual-level user-aware fine-tuning. We first train an intermediate model to understand fun at a population level, then use it to serve as a better initialized point for downstream user-aware fine-tuning. As we will show, the intermediate model is already good at making zero-shot inference. Further, it makes adapting to each individual's unique perception of fun much more accurate and stabler, in a few-shot fine-tuning setting.

### 3 Intermediate Task: Can Language Models Understand What Fun is?

Under the current pre-training - fine-tuning paradigm in solving NLP tasks, we expect further an intermediate task (Poth et al., 2021) to bridge a vanilla pre-trained language model to a user with a few labeled data points that indicate their unique perception toward fun. To this end, an intermediate language model that has been fine-tuned to understand what fun generally is at a **population level** is needed. Such of an intermediate model allows faster and stabler adaptation to specific users in downstream applications as we will show in Section 4.

### 3.1 Dataset Collection

We aim to look for a one-size-fits-all data source that is targeted toward readers in seek of fun, enabling their reactions to be used as a proxy indicator of this concept; while covering as diverse genres as possible, serving as an inclusive cornucopia of human language to learn a generic model for domain-agnostic perception of fun, instead of focusing on a specific domain/topic. Extracting data from one source eliminates cross-dataset annotation inconsistency.

**Example Data Source** In this work, we realize these above-mentioned heuristics by presenting an effective data source. We highlight that this is an example data source that can demonstrate the utility of our designed mechanism and as proof of concept, yet not an optimal one. Cracked.com is based on the Cracked magazine, which collects interesting content covering topics from movies, TV, video games, music, sports, history, science, sex, tech, news, celebrities, to "weird world". Albeit it claims to be "the America's only humor site", we find that the linguistic connotation of the spirit behind the platform is extremely close to our defined concept of *fun*, as the site covers substantial informative content from a wide range of fields, *instead of the well-perceived definition of humor*, which mostly concerns punchlines and jokes (Mihalcea and Strapparava, 2005; Yang et al., 2015).

### 3.2 Automatic Scoring Mechanism

We make the intuitive assumption that readers of the data source are a specific group of people seeking for fun content. Thus, simple features like # of comments are intuitively highly correlated with the measurement of **engagement**. Further, it is in a **liberating** fashion due to the nature of the platform studied. Therefore, readers' reactions approximate how fun the content is - how much **liberating engagement** is shown at a population level.

We scraped all the posts from the website, from Jan, 2005 to March, 2022, yielding over 15k articles on diverse topics. In the same vein as Yang et al. (2021) who used naturally available user reactions on twitter posts as a proxy to indicate humor, we novelly define an automatic scoring mechanism to annotate how *fun* the content is as follows:

$$Fun = \tanh\left(\frac{n}{\alpha \cdot \frac{1}{|I_y|} \sum_{i \in I_y} n_i}\right), \quad (1)$$

where  $n$  denotes the number of comments in

an observed article;  $I_y$  denotes the set of articles in the specific year that the article falls in, where  $I_y \in I$ , with  $I$  representing the full set of articles from 2005 to 2022. We perform a mean normalization over each year by dividing the number of comments by the sum of all  $n_i$  in a given year  $I_y$ , since the posts show significantly different average number of comments every year. Generally, newer articles get less comments to date. Intuitively, our defined metric serves as a proxy indicator of the engagement of a post compared to other posts in the same year. We further introduce a coefficient  $\alpha$  to denote "non-tolerance toward not fun", which together with  $\tanh()$ , could be used to twist the distribution of fun scores (Figure 2). Mathematically, a higher  $\alpha$  (higher non-tolerance toward not fun) makes the score distribution right-skewed, as less posts could receive high fun scores. Notably,  $\alpha$  can potentially be parameterized in later user-aware fine-tuning of downstream tasks (giving each user an interpretable non-tolerance score), which is out of scope of this paper.

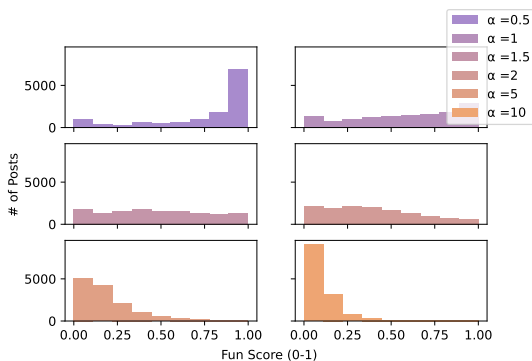


Figure 2: Twisting fun score distribution with  $\alpha$

### 3.3 In-domain Learning of Fun

To validate the proposed scoring mechanism and the utility of *fun* as a universal metric to understand content in a continuous space, we first conduct an experiment with three language models: BERT-base (Devlin et al., 2019), RoBERTa-base (Liu et al., 2019), and LongFormer-base (Beltagy et al., 2020). 10% of the data is used as the hold-out test set. The models are fine-tuned with over 50 epochs (Zhang et al., 2020) for stabler performance, with a learning rate of  $3e - 5$ , instead of the commonly adopted 3 epochs.

The result is shown in Table 1. We observe that, while early epochs are extremely fluctuating - given the regression nature of the task and the nuanced

Model	F1	R-Squared
BERT <sub>base</sub>		
+ max seq. 128	0.748	.437
+ max seq. 512	0.751	.460
RoBERTa <sub>base</sub>		
+ max seq. 128	0.753	.448
+ max seq. 512	0.758	<b>.499</b>
LongFormer <sub>base</sub>		
+ max seq. 2048	<b>0.760</b>	.497

Table 1: In-domain learning

nature of the concept studied, the models can typically reach to an optima where increased epochs bring diminished fluctuations. In the later epochs of training, RoBERTa can stably converge to a performance close to .50 R-squared and an F1 score over 75% on the test set. To attain an F1 score, we transform the 0-1 scale in the regression task into a binary classification using a decision threshold (0.5 by default). Given how "fun" is less of a universal phenomena compared to "humor", our experiment demonstrates a surprising result - achieving a performance even better than what Weller and Seppi (2019) reported on a similar task using Reddit joke posts and their up-votes, which again in turn proves the utility of our scoring mechanism to be learned against by the LMs.

**Length matters for fun** A linguistic property of fun we try to validate is the importance of informative content conveyed, which we hypothesize to be supported by the length of the text, distinguishing the concept of *fun* from existing studies on *humor*, with the latter mostly characterized as one-liners (Mihalcea and Strapparava, 2005; Yang et al., 2015) that typically have a maximum sequence length of 10-30 words. We run both BERT and RoBERTa with a maximum sequence length of 128 and 512 tokens. This part of the experiment shows that length does matter in order for textual content to be perceived as "fun", corroborated by the R-squared performance boost by a margin of respectively 5.3% and 11.4% on BERT and RoBERTa brought by taking in a longer text. This is further validated by the on-par performance of Longformer, which can be thought of as a variant of RoBERTa that enables computational complexity to scale linearly with sequence length. For Longformer, we consider a maximum of 2048 tokens. Notably, taking in texts that are too long brings less pronounced advantages, which we hypothesize to be due to:

(1) the limited attention of readers in seek of "fun" distribute to in-depth reading, and (2) there exists a certain length that suffices for readers to engage - 512 tokens suffice in our case.

## 4 Helping Individual Users to Break through Inequality in Information Acquisition

When deploying in a user-specific setting, we expect the in-domain training in section 3.3 to serve as an intermediate task (Poth et al., 2021). It is thus of interest to study how well the previous intermediate pre-training on the in-domain data source could transfer to users, respectively under (1) zero-shot inference and (2) further fine-tuning settings.

### 4.1 Zero-shot Inference

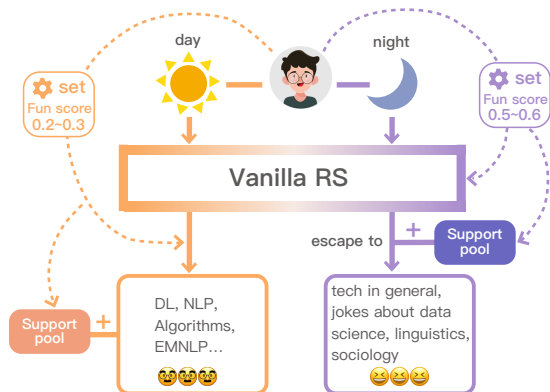


Figure 3: User case: "A day of an NLP researcher", empowered by our 'Fun' framework. While trapped in the cocoons of NLP-related content, one is given the option to turn up the 'Fun' range to receive recommendations from other domains, generated from the vanilla content pool and borrowed from a supporting pool.

Complementing the example that "an NLP researcher might not be keen to read about NLP before they go to bed" (Figure 3), we present an interesting case study.

We consider a binary setting, by taking in two target data sources, Medium<sup>2</sup> and Not Always Right<sup>3</sup>. We first collect 50 posts from Medium, using an account that has been "fine-tuned" by an NLP researcher by using the account for a year. The recommended posts are all about NLP. We then collect the same number of posts from Not Always Right, a website dedicated to high-quality fun stories.

<sup>2</sup> <https://medium.com/>

<sup>3</sup> <https://notalwaysright.com/>

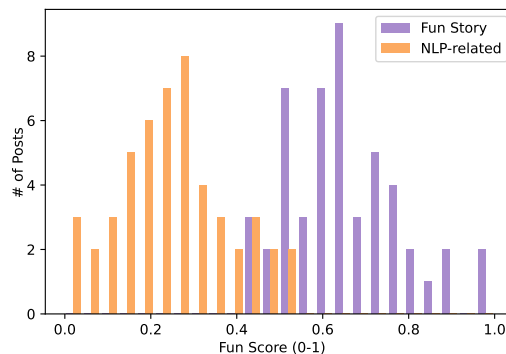


Figure 4: Significant distribution discrepancy in zero-shot cross-domain inference, using RoBERTa checkpoint trained in Section 3.3

We then leverage the model checkpoints trained to understand *fun* in section 3.3 to infer on them. Figure 4 shows the predictions yielded on these two target sources, indicating a significant distributional discrepancy ( $p < .001$ ). This further proves that in an industrial deployed setting, it is possible to help a specific group of people escape to content of other genres and domains, by sampling posts of user-determined fun score from a supporting set. We further propose a possible way to implement this framework in Appendix A, through a plug-in user interface we design.

### 4.2 Cross-Domain User-aware Fine-tuning

In this section, we study with three human annotators, with the goal of adapting a language model to their unique perception of fun. We ask them to annotate 300 posts each from a generic Medium dataset covering a wide range of genres and topics. With the annotated dataset we aim to find the best strategy to adapt to their exact perception, under this few-shot learning setting (210 posts excluded 30% data as test set). We find that the in-domain training in Section 3.3 serves as a powerful intermediate task that stabilizes the following user-aware further fine-tuning.

#### 4.2.1 User Annotation

**Dataset** To perform experiments on cross-domain fine-tuning, we leverage a 190k Medium dataset<sup>4</sup>. As opposed to directly collecting Medium posts on an account "fine-tuned" by a user in Section 4.1, this generic Medium dataset comprises of high-quality articles of diverse topics and genres.

<sup>4</sup> <https://www.kaggle.com/datasets/fabiochiusano/medium-articles>


Fun Spectrum	Description and Scoring Guidance
	$0.8 < \text{fun score} \leq 1.0$ : Make you forget surroundings and have inner motives to practice what you see from it.
	$0.6 < \text{fun score} \leq 0.8$ : Feel intrigued and devote a great amount of attention to repeat watching or reading it.
	$0.4 < \text{fun score} \leq 0.6$ : Start to have revealing happiness (e.g., laughter) and is likely to share or imitate the content.
	$0.2 < \text{fun score} \leq 0.4$ : Learn something new or inspiring but require self-supervision to keep concentrated.
	$0 < \text{fun score} \leq 0.2$ : Not very interested and reluctant to pay attention to if needed.
	0: Feel detested or indifferent, not willing to spend time on.

Table 2: Annotation Instruction for Cross-domain Fine-tuning

We ask 3 annotators to annotate their fun perception on 300 articles each, based on the annotation instruction in the next sub-section. To validate the **non-universality** of fun perception of different individuals, we secretly put a same pool of 100 articles as the last 100 posts they annotate. Other 200 articles for each annotator are randomly sampled from the 190k articles.

**Annotation Guidelines** Given the cruel fact that creating an unequivocal definition of "fun" is almost impossible, for its discursive and subjective nature, we design a reference instruction (Table 2) for target groups to annotate fun score.

From game designing to physical education and leisure studies, fun has gained increasing attention and revealed some enlightening commonalities: positive emotion and sense of engagement. When exposed to something fun, some may laugh, while others are likely to share. Whatever their actual actions are, these fun moments are engaging and getting people focused. This illustrates the essential characteristics of fun: a broader ideation encompassing not only punchlines, but educational and inspiring content that may have further influences on individuals, distinguishing it from humor and funny.

Thus, knowing "how fun manifests in real life" is what really matters. Correspondingly, based on previous studies, qualitative descriptions are used in the instruction sheet as presentation of fun. We expect respondents to annotate provided content, referring to the sectional instruction (Table 2), and give out their fun perception on a specific post. Scoring could be more granular than the instructed range (e.g., 0.66) to accord with the diffused feature of fun in real life.

**Non-universality of Fun Perception** We further present the inter-annotator agreement in Table 3, computed on the secret pool of 100 articles shared

Inter-Annotator Agreement	
Pearson's r	
User 1 and 2	0.536
User 1 and 3	0.457
User 2 and 3	<b>0.652</b>
Krippendorff's alpha	
All users	0.405

Table 3: Inter-Annotator Agreement

by the three annotators. It is shown that users have unique perception toward fun. For instance, user 1 shares lower Pearson's correlation with the other two annotators. On top of that, the Krippendorff's alpha among the three annotators is not extremely high, again indicating the non-universality of fun perception. However, even for the most sophisticated individual, our method yields better results (Table 4 as described in the 4.2.2) than the baseline methods, showing the performance boost brought by the intermediate pre-training and as well proving the utility of our automatic scoring mechanism to learn against. However, the sophisticated nature of some users' understanding toward fun in turn necessitates more training data from them.

#### 4.2.2 User-aware Fine-tuning Performance

In this section, we leverage the checkpoints in in-domain training to learn further in a user-aware setting.

Table 4 shows the results of different fine-tuning settings, and the superiority of our methods. For simplicity, we only present experiments with RoBERTa (Liu et al., 2019), since it provides the best result (Section 3.3) and is computationally cheaper than Longformer (Beltagy et al., 2020) if the latter takes in a maximum of 2048 tokens.

For the baseline method, we directly take a RoBERTa pre-trained checkpoint, and fine-tune it on the user data. For our methods, we define

Model	Avg. R-squared	Max. R-squared	Min. R-squared	Var.
<i>User 1</i>				
RoBERTa + ft.	.016	.256	-.624	.019
RoBERTa <sub>fun-shallow</sub> (ours) + ft.	.053	.248	-.356	.013
RoBERTa <sub>fun-deep</sub> (ours) + ft.	<b>.093</b>	<b>.265</b>	<b>-.138</b>	<b>.010</b>
<i>User 2</i>				
RoBERTa + ft.	.274	.544	-.706	.063
RoBERTa <sub>fun-shallow</sub> (ours) + ft.	.412	.626	-.081	.025
RoBERTa <sub>fun-deep</sub> (ours) + ft.	<b>.450</b>	<b>.660</b>	<b>-.003</b>	<b>.021</b>
<i>User 3</i>				
RoBERTa + ft.	.085	.519	-1.551	.225
RoBERTa <sub>fun-shallow</sub> (ours) + ft.	.208	.519	-2.816	.191
RoBERTa <sub>fun-deep</sub> (ours) + ft.	<b>.407</b>	<b>.520</b>	<b>-.105</b>	<b>.011</b>

Table 4: Cross-domain fine-tuning performance

a RoBERTa<sub>fun-shallow</sub> and a RoBERTa<sub>fun-deep</sub>. For RoBERTa<sub>fun-shallow</sub>, we first fine-tune a RoBERTa checkpoint on our in-domain Cracked.com dataset for 3 epochs, then further fine-tune it on the user-annotated data. While for RoBERTa<sub>fun-deep</sub>, we do the same for 50 epochs, then further fine-tune. The utility of pre-training long enough on our in-domain dataset is proved as shown in Table 4.

We hypothesize that, this amount of user data is not sufficient for a pre-trained language model to understand **what is fun** from scratch. A language model is not able to accurately capture which words and what combinations of them make an article fun through a few hundred of examples of diverse genres. Even though some articles show similar topics and fun scores, it is extremely hard for a RoBERTa to find the salient areas to put attention to through a few examples, if it is allowed to look at the first 512 tokens of an article.

By contrast, our methods mitigate this inefficiency through providing an extra resource of 15k articles to indicate what fun is. Albeit being "cross-domain", words, expressions and language in general that express the concept of **fun** could be quite transferable. As shown in the result, training long enough (50 epochs) first on our in-domain dataset enables the models to attain deeper understanding of fun, for later stabler transferring.

For each run, we hold out 30% data (90 articles) as test set for each user. For each setting and user, we repeat each further fine-tuning 3 times with 50 epochs, with different random seeds and data splits, yielding 150 unique stages for each combination, to get a closer look at the training progress (The results in Table 4 are based on these epoch-

level computations). We again use R-squared as the evaluation metric which measures how well a regression model explains the observed data.

The baseline method is extremely fluctuating throughout the 50 epochs for each run, occasionally "hitting" a not-bad result on the test set and could drop dramatically in the next epoch, resulting in extremely high variance. By contrast, our methods typically "wander" around at a stabler range of R-squared on the test set throughout the training, with significantly lower intra-epoch variance. In real-life deployment, this stability is extremely important in efficiently adapting to user preferences without drastically fluctuating performances. Notably, the baseline method typically reach and wander around an average training loss of 0.004 in the last few epochs, while our methods lead it to an average training loss of 0.0005, meaning that the pre-training on the in-domain dataset has already put the training to a better optima. Thus, it is evident that giving more user-specific data, the further fine-tuning could be more robust with our in-domain intermediate training.

Also, we give out a description of attaining user-specific data and user-aware fine-tuning in a possible real-life deployment at the end of Appendix A, on top of the vanilla zero-shot intermediate models described in the last section.

## 5 Conclusion

In this work, we identify a misalignment of a (content, concept) pair existing among social classes - one's information cocoon and the corresponding degree of fun. This passively decided characteris-



tic might reinforce, as opposed to help deconstruct the traditional social structure of beneficiaries, and therefore is originally not a manifestation of the advancement of AI algorithms contributing to social good.

In combat with this prevalent issue, we propose an NLP framework, which is composed of 1) an intermediate language model that could understand/predict the degree of liberating engagement (the degree of fun) of text at a population level. 2) further user-aware fine-tuning method to adapt the intermediate model to each individual's unique perception of fun. Moreover, we propose some possible real-life cases to deploy our framework in a platform-agnostic setup as an external regularization over recommender systems, such as a web-based plug-in that could filter content based on user-defined fun range, without having to explicitly interact with or adjust the algorithms behind the recommender system of a platform.

## Limitations

We consider our work to be a framework that mitigates the inequality in the traditional social structure of beneficiaries (Jin et al., 2021), and establish it to be a work promoting "NLP for social good". Nonetheless, we envision more detailed implementations to be studied that regulate its usage in practice. As Jin et al. (2021) put it, stage-4 NLP applications should be most careful about their ethical concerns. Our work can also be posed with extreme use cases. For instance, what if our framework in turn helps under-privileged people to immerse themselves in entertainment content? - if they are given the complete freedom to do so. Therefore, we call for more regulated usage of our proposed framework. However, we position our framework to be one that provides humans with dignity, freedom, and fairness (Zeng et al., 2019), under an era where discriminating AI algorithms are prevalent.

Moreover, we do notice some biases existing in the in-domain *fun* models, brought by domain-specific content in our scraped dataset (Cracked.com). For instance, inferring with these models on a one-word input, "LGBT", gives a score of around 0.87-0.95, showing the extensive interest that American people hold for politically correct-related discussions on controversial issues. With "sequence length is a domain" (Varis and Bojar, 2021) considered, these biases are significant.

Thus, we highlight that what we present here is a novel framework, rather than an optimized solution. The dataset we leverage is a demonstration of a feasible way of empowering this framework, rather than an optimal one. We envision more endeavors to be made for more generic annotated data for this concept. Although in this paper we provide what can be formulated as a "distant-supervised approach" (*i.e.* an automatic scoring scheme) for the sake of studying this concept, annotation can still potentially provide less bias toward learning on the concept of *fun* that is loyal and inclusive (Joshi et al., 2020) to the perception of a wider audience from all socio-demographic groups.

## Acknowledgements

We thank Haoting Dai from Princeton University, Hanshen Li from Alibaba-Ant Group, Zhongtian Sun, Tom Winterbottom and James Burton from Durham University for their helpful discussion and valuable feedback. We thank three anonymous reviewers for their valuable comments.

## References

- Pablo Barberá, John T Jost, Jonathan Nagler, Joshua A Tucker, and Richard Bonneau. 2015. Tweeting from left to right: Is online political communication more than an echo chamber? *Psychological science*, 26(10):1531–1542.
- Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.
- Enrique García Bengoechea, William B Streaan, and DJ Williams. 2004. Understanding and promoting fun in youth sport: coaches' perspectives. *Physical Education & Sport Pedagogy*, 9(2):197–214.
- Mark Blythe and Marc Hassenzahl. 2018. The semantics of fun: Differentiating enjoyable experiences. In *Funology 2*, pages 375–387. Springer.
- Mark A Blythe, Kees Overbeeke, Andrew F Monk, and Peter C Wright. 2004. *Funology: from usability to enjoyment*. Springer.
- Charles H Cooley. 1955. Primary groups. *Small groups*, pages 15–17.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages

- 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Hsiao-tung Fei, Xiaotong Fei, Gary G Hamilton, and Wang Zheng. 1992. *From the soil: The foundations of Chinese society*. Univ of California Press.
- Ben Fincham. 2016. *The sociology of fun*. Springer.
- Bryan T Gervais. 2015. Incivility online: Affective and behavioral reactions to uncivil political posts in a web-based experiment. *Journal of Information Technology & Politics*, 12(2):167–185.
- Erving Goffman. 1961. *Encounters: Two studies in the sociology of interaction*. Ravenio Books.
- Mark Granovetter. 2018. *Getting a job: A study of contacts and careers*. University of Chicago press.
- Mark S Granovetter. 1973. The strength of weak ties. *American journal of sociology*, 78(6):1360–1380.
- Jaron Harambam, Natali Helberger, and Joris van Hoboken. 2018. Democratizing algorithmic news recommenders: how to materialize voice in a technologically saturated media ecosystem. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 376(2133):20180088.
- Julie Jiang, Xiang Ren, Emilio Ferrara, et al. 2021. Social media polarization and echo chambers in the context of covid-19: Case study. *JMIRx med*, 2(3):e29570.
- Zhijing Jin, Geeticka Chauhan, Brian Tse, Mrinmaya Sachan, and Rada Mihalcea. 2021. [How good is NLP? a sober look at NLP tasks through the lens of social impact](#). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3099–3113, Online. Association for Computational Linguistics.
- Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. [The state and fate of linguistic diversity and inclusion in the NLP world](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
- A MacPhail, T Gorely, D Kirk, and G Kinchin. 2008. Exploring the meaning of fun in physical education through sport education. *Research Quarterly for Exercise and Sport*, 79(13):344–356.
- Rod A Martin and Thomas Ford. 2018. *The psychology of humor: An integrative approach*. Academic press.
- Rada Mihalcea and Carlo Strapparava. 2005. [Making computers laugh: Investigations in automatic humor recognition](#). In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 531–538, Vancouver, British Columbia, Canada. Association for Computational Linguistics.
- Travis Tae Oh and Michel Tuan Pham. 2022. A liberating-engagement theory of consumer fun. *Journal of Consumer Research*, 49(1):46–73.
- Walter Podilchak. 1991. Distinctions of fun, enjoyment and leisure. *Leisure studies*, 10(2):133–148.
- Clifton Poth, Jonas Pfeiffer, Andreas Rücklé, and Iryna Gurevych. 2021. [What to pre-train on? Efficient intermediate task selection](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10585–10605, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Nils Reimers and Iryna Gurevych. 2019. [Sentence-BERT: Sentence embeddings using Siamese BERT-networks](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.
- Daniel Röchert, Gautam Kishore Shahi, German Neubaum, Björn Ross, and Stefan Stieglitz. 2021. The networked context of covid-19 misinformation: Informational homogeneity on youtube at the beginning of the pandemic. *Online Social Networks and Media*, 26:100164.
- Willibald Ruch. 2001. The perception of humor. In *Emotions, qualia, and consciousness*, pages 410–425. World Scientific.
- Tara K Scanlan and Jeffery P Simons. 1992. The construct of sport enjoyment. *Motivation in sport and exercise*, 1992:15.
- Shaoyun Shi, Weizhi Ma, Zhen Wang, Min Zhang, Kun Fang, Jingfang Xu, Yiqun Liu, and Shaoping Ma. 2021. Wg4rec: Modeling textual content with word graph for news recommendation. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 1651–1660.
- Yu Song, Shuai Sun, Jianxun Lian, Hong Huang, Yu Li, Hai Jin, and Xing Xie. 2022. Show me the whole world: Towards entire item space exploration for interactive personalized recommendations. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, pages 947–956.
- Cass R Sunstein. 2006. *Infotopia: How many minds produce knowledge*. Oxford University Press.
- Cass R. Sunstein. 2007. *Republic.com 2.0*. Princeton University Press.

- Dusan Varis and Ondřej Bojar. 2021. [Sequence length is a domain: Length-based overfitting in transformer models](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8246–8257, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Ning Wang. 1999. Rethinking authenticity in tourism experience. *Annals of tourism research*, 26(2):349–370.
- Orion Weller and Kevin Seppi. 2019. [Humor detection: A transformer gets the last laugh](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3621–3625, Hong Kong, China. Association for Computational Linguistics.
- Huimin Xu, Zhicong Chen, Ruiqi Li, and Cheng-Jun Wang. 2020. The geometry of information cocoon: Analyzing the cultural space with word embedding models. *arXiv preprint arXiv:2007.10083*.
- Diyi Yang, Alon Lavie, Chris Dyer, and Eduard Hovy. 2015. [Humor recognition and humor anchor extraction](#). In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2367–2376, Lisbon, Portugal. Association for Computational Linguistics.
- Zixiaofan Yang, Shayan Hooshmand, and Julia Hirschberg. 2021. [CHoRaL: Collecting humor reaction labels from millions of social media users](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4429–4435, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Yi Zeng, Enmeng Lu, and Cunqing Huangfu. 2019. Linking artificial intelligence principles. In *SafeAI@AAAI*.
- Tianyi Zhang, Felix Wu, Arzoo Katiyar, Kilian Q Weinberger, and Yoav Artzi. 2020. Revisiting few-sample bert fine-tuning. In *International Conference on Learning Representations*.

## A Industrial Deployment: A Proposal

We also demo a possible case to deploy our framework in real-life scenarios. We design a Chrome extension (Figure 5) that can filter existing recommended posts based on user-defined fun range. A Chrome extension allows us to retrieve information from the platform pages, as well as injecting HTML tags to them. We propose it to be a simple way to deploy our model as an extension backend to quantitatively score the text data that is captured,

and is generalizable across different platforms without having to explicitly interact with their recommender systems.

Figure 6 shows the working process of the Chrome extension we develop for Quora.com. After users install the *fun* extension on Chrome and start it, the extension would start to continuously capture posts under the current page and check whether the current posts have been scored according to their IDs in the database. If a post ID has not existed, the text content of that post will be input to our model for scoring, and the score with the corresponding post ID will be stored in the Web SQL Database and display to the bottom of that post on the page.

Figure 5 shows our framework deployed as an plug-in that can be applied to Quora.com. As an example, when the user selects fun scores between 0.6 and 0.8 and clicks the **Search** button, the current page will be filtered based on the selected range and display the first 50 posts within that range.

Moreover, it is possible for users to modify the fun scores inferred by the vanilla zero-shot model. These modified fun scores would be used as ground-true user-specific fun perception to fine-tune a user-aware model, making it possible to understand user’s unique perception for better "escape".

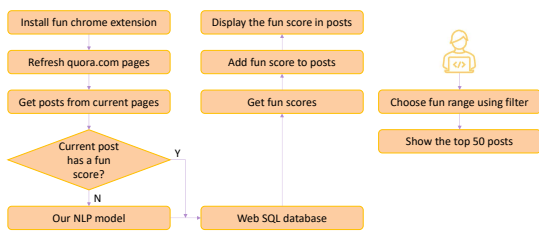


Figure 6: Fun chrome extension workflow

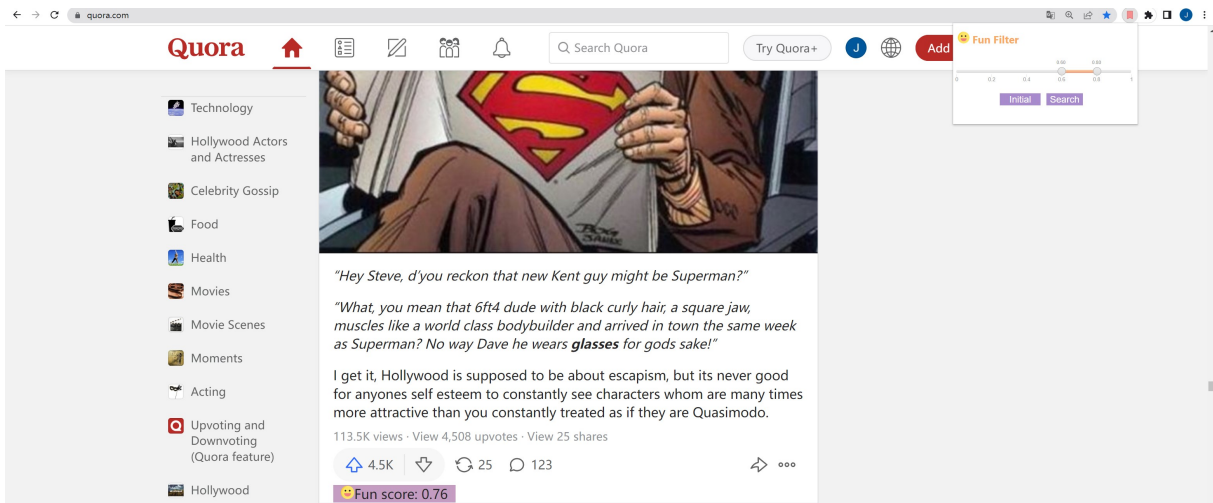


Figure 5: Deployment of fun filter in quora.com