Blind Leading the Blind: A Social-Media Analysis of the Tech Industry

Tanishq Chaudhary* **Pulak Malhotra*** **Radhika Mamidi** **Ponnurangam Kumaraguru**
IIIT Hyderabad
tanishq.chaudhary@research.iiit.ac.in, pulak.malhotra@students.iiit.ac.in ,
{radhika.mamidi, pk.guru}@iiit.ac.in

Abstract

Online social networks (OSNs) have changed the way we perceive careers. A standard screening process for employees now involves profile checks on LinkedIn, X, and other platforms, with any negative opinions scrutinized. Blind, an anonymous social networking platform, aims to satisfy this growing need for taboo workplace discourse. In this paper, for the first time, we present a large-scale empirical text-based analysis of the Blind platform. We acquire and release two novel datasets: 63k *Blind Company Reviews* and 767k *Blind Posts*, containing over seven years of industry data. Using these, we analyze the Blind network, study drivers of engagement, and obtain insights into the last eventful years, preceding, during, and post-COVID-19, accounting for the modern phenomena of work-from-home, return-to-office, and the layoffs surrounding the crisis. Finally, we leverage the unique richness of the Blind content and propose a novel content classification pipeline to automatically retrieve and annotate relevant career and industry content across other platforms. We achieve an accuracy of 99.25% for filtering out relevant content, 78.41% for fine-grained annotation, and 98.29% for opinion mining, demonstrating the high practicality of our software.

1 Introduction

In the last two decades, online social networks (OSNs) like Facebook, Instagram, LinkedIn, and X have dramatically impacted our lives. By converting our limited offline social capabilities into digital form, OSNs enable exponential reach, touching all aspects of life. One such area with a colossal impact is our careers.

Online profiles on the aforesaid platforms are now used to screen and terminate employees. History and activity are examined and scrutinized, and any negativity is marked (Aichner et al., 2021). While positive discussions thrive, the other side of the story is stifled. Critical discourse on already taboo topics – toxic culture, poor management, and especially fair compensation – are thus completely left out of the picture.

Strong reasons for honesty gave rise to the platform TeamBlind (often just shortened to Blind), with the first post in 2015. Since then, Blind has grown to 9M+ employees from 300k+ companies.[1]

There are two characteristics of Blind that make it valuable. First, Blind is entirely anonymous. Users are free to post negative and positive reviews of companies, freely listing the pros and cons of their workplace. At the heart of Blind lies the discussion section, allowing for productive discussions on any topic whatsoever – be it HR issues, career-related, broader industry-related, referrals, layoffs, and more. People are free to comment underneath and provide contrasting opinions – an idea non-existent in the previous platforms. Second, Blind has a strict verification check. Without a login to the network, a user can only view two posts for free. To participate in the discussions and leave reviews, the user must log in via a work email. This step is strict, and it was (and is) crucial to Blind's continued growth. Due to a lack of strict checks, another platform with similar motives, Glassdoor, got poisoned by the epidemic of fake reviews.[2]

Blind has invited both supporters and critics, with the former arguing for the wealth and variety of opinions and the latter citing some level of toxicity that anonymity brings. Without a doubt, both agree that the platform has reshaped how we discuss careers. Hence, a detailed analysis of the Blind platform will shed some light on the pulse of the

*These authors contributed equally to this work

[1] Sources vary on the exact numbers. We consider the latest numbers from the official website, https://www.teamblind.com/.

[2] https://www.gadgetreview.com/fake-reviews-glassdoor

industry.

2 Blind Data

Blind has three main ways in which users interact with the platform: *Blind Company Reviews*, *Blind Posts*, and company-specific arenas. We examine the first two since they provide broader, across-company analyses. All the data is collected from Blind's inception on October 20, 2015, until February 2023 – totaling approximately seven years of data. We release the two following novel datasets using FAIR principles.[3]

We follow strict ethical guidelines by excluding any personally identifiable information, like the names of users or any details that could lead to identification.

2.1 Blind Company Reviews

Users on Blind can review their current or past companies, where each company review consists of a list of textual pros and cons and ratings out of five for various metrics (rating of one being the worst and five being the best). The first is an overall rating for a company, followed by career growth, management, culture, work-life balance, and compensation. We scraped the website using Selenium and gathered **63,477** company reviews from the 55 most reviewed companies on Blind. We call this dataset *Blind Company Reviews*.

2.1.1 Content Analysis

Focusing on the contents of the pros and cons field, we found that users are more verbose in listing their company's cons rather than the pros (Figure 1). A t-test showed statistical significance (p-value < 0.05) for 39 out of 55 companies. These results indicate the negativity bias phenomenon, which parallels findings in cognitive and social psychology (Kanouse et al., 1972). Only five companies show an inversion of verbosity (with p-value > 0.05).

To better understand the content of the pros and cons, we analyze the frequencies of unigrams and bigrams. Jointly, we found that work-life balance (WLB) has the most mentions, coming up as "work life" and "life balance" in bigrams. For the former, people talk about "good compensation", "great work", "great culture", "smart people", and "great benefits". Latter's analysis shows discussions of similar topics in an antagonistic light, for example,

[3] https://www.go-fair.org/fair-principles/

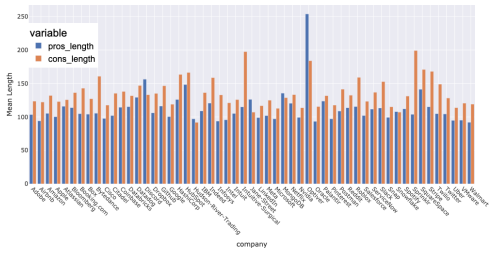


Figure 1: Company review lengths for pros and cons by number of characters.

"slow career", "bad work", and "low compensation". We also observe cons related to poor management described with adjectives like "middle", "bad" and "upper", hindrance of decision-making in large companies, "red tape", and old and unmaintained code, "tech stack".

These findings strongly validate our data, demonstrating the unique dual discourse of Blind by including topics otherwise left as workplace taboos.

2.1.2 Metrical Analysis

Now, we turn to the star ratings on different metrics in a review. Most rate the company 4 or 5 stars (overall rating), with minimal 1-star and 2-star reviews. This could mean the employees are mostly satisfied with their employer, even as they elaborate more on a company's negatives.

We find surprising significant correlations (p-value < 0.001) using Spearman's r across all companies between all the ratings (Figure 2). We expect the work-life balance to be highly correlated with the company ratings since users talk about it the most (Section 2.1.1). As it turns out, the ratings are instead most correlated with culture (r=0.71) and management (r=0.7) and the least correlated with WLB (r=0.49). This might mean poor management and toxic cultures are the biggest consistent reasons for a lower rating. Since WLB has the lowest r, we expect there to be cases with low WLB but high overall ratings.

We explore the causes for this anomaly by finding the top three companies by mean rating for each metric (Table 1). The median is 5 for all the entries. High-frequency trading (HFT) companies Hudson-River-Trading (HRT), Jane Street (JS), and Optiver are consistently at the top. These firms are known to have long and stressful hours – excellent compensation but low WLB, which explains the low r value for WLB. Only the WLB column shows a different composition of companies, with Indeed,

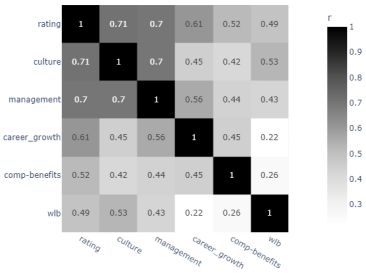


Figure 2: Correlation matrix of ratings across companies.

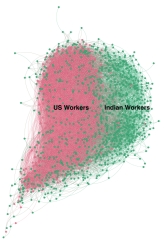


Figure 3: Cluster based on user interaction in the USA (pink nodes) and India (green nodes).

SquareSpace, and Atlassian being the best, while the HFTs are not found even in the top ten.

2.2 Blind Posts

The most treasured part of the Blind platform is the posts section. Any logged-in user can create an unstructured text post and assign a relevant board name (post category). Others can view, like, and comment underneath with a depth of 2 – comments of comments are allowed, but further comments are not. Using Blind's public API, we collect **767,224 posts** from 74 different boards.

2.2.1 Preliminary Analysis

Most posts on Blind are below 500 characters and 80 to 120 words. N-gram study of the posts shows "total compensation" as the most frequent bigram by far, with mentions in 193k posts – a symptom of Blind's obsession with transparent pay.

MAMAA companies dominate the unigram rankings in content and hashtags alike.[4] Furthermore, out of 83.03% posts that mention the poster's company name, only 20 companies (< 0.0067%) make up for 54% of the posts. All of these companies are either big tech or tech-related. Even though Blind is meant to be a platform for employees of any industry, we see a remarkable gravity in the tech industry. We explore and exploit this in the following sections.

3 Network Characterization

In this section, we zoom out and gain a broader perspective on the entire Blind platform. We look at the platform's interactions, quantifying the drivers of user engagement and how COVID-19 has impacted the tech industry.

[4]MAMAA is commonly used acronym used for Meta, Apple, Microsoft, Amazon, and Alphabet

3.1 Network Graph

For the posts that mention "total compensation", 85.48% are in USD, 10.97% are in INR, and the remaining are spread across multiple currencies. Leveraging the currency as the proxy for the country, we explore the interactions between users in the USA and India, the two big tech hubs (Figure 3).[5] We represent users as the nodes (pink for the USA and green for India) and interactions between the users as edges – linked through comments on the same post or a comment thread. Due to the high computational complexity of this operation, we sample only the most active 1000 users from either country.

Interestingly, instead of a unified mixed cluster, two distinctive communities emerge. Users from the USA make up little more than half of the platform activity. We also observe a lesser number of Americans participating in the Indian discussions as compared to the Indians on the American side. This difference may be attributed to the outflow of Indians abroad for work, with Indians taking more interest in discussions everywhere.[6]

3.2 User Engagement

We measure the engagement of a post by the number of views, likes, and comments it receives. The number of views shows a log-normal distribution, following Gibrat's rule of proportionate growth (Mansfield, 1962). This means that popular posts gain more traction exponentially quickly. Unique to the platform, we observe 14.03 ± 42.754 comments per post – while the number of likes is unexpectedly lower, 3.63 ± 30.287.[7] This is in contrast

[5]https://hbr.org/2023/04/the-u-s-india-relationship-is-key-to-the-future-of-tech

[6]https://www.fortuneindia.com/macro/more-indians-plan-on-moving-abroad-in-next-2-yrs-survey/111674

[7]Note that the likes and comments do not follow the normal distribution, which explains lower than zero values for one standard deviation away from the mean.

Overall	Culture	Management	Growth	WLB	Comp
HRT (4.76,0.23)	HRT (4.79,0.26)	HRT (4.47,0.49)	Optiver (4.64,0.53)	Indeed (4.58,0.67)	Optiver (4.90,0.08)
JS (4.73,0.40)	Discord (4.61,0.72)	JS (4.45,0.79)	JS (4.53,0.46)	SquareSpace (4.57,0.54)	HRT (4.88,0.15)
Optiver (4.63,0.53)	JS (4.61,0.61)	Optiver (4.14,1.36)	HRT (4.30,0.79)	Atlassian (4.56,0.66)	JS (4.84,0.18)

Table 1: Top three companies for the rating metrics. The mean and variance for each are mentioned respectively.

to other platforms (like Reddit (Baumgartner et al., 2020), for example), where people find it more convenient to like a post and move on. Apart from the high variance, we hypothesize this disparity is due to the platform's anonymity, which generates productive discussions in comment threads rather than passively liking a post. We find significant correlations (p-value = 0) between views and likes (Pearsons' r=0.551), lesser than the coefficient for views and comments (r=0.746).

3.3 Temporal Network Activity

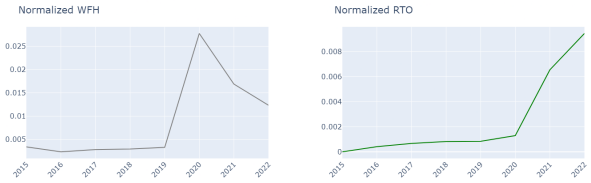
To understand the Blind platform's characteristics across the years, we consider all *Blind Posts* made on the platform since its start in 2015. We find the Year-Over-Year activity for the posts (Figure 6), which is paralleled by the number of reviews posted (Figure 6, inset). Considering that most of the platform is dominated by tech-industry employees, our analysis can shed light on the peaks and valleys of the tech industry. Specifically, we consider three critical events alongside the COVID-19 pandemic: the advent of the work-from-home (WFH) options, return-to-office (RTO), and the recent layoffs that have led to more than 200k+ job cuts.[8] For each event, we assign binary labels based on the existence of keywords – "work from home" (or "wfh"), "return to office" (or "rto"), and "layoff" respectively, and normalize the aggregated score for a year by its activity.

3.3.1 Work From Home

We observe negligible mentions for WFH till 2019 (Figure 4a). WFH peaked in 2020 when the pandemic spread and has since declined as firms call back employees (Bick et al., 2020).

3.3.2 Return To Office

Compared to WFH, an opposite trend is observed for return-to-office (RTO) (Figure 4b). The RTO trend starts one year later, in 2021, after WFH peaks. This results from nations starting to open back up after the pandemic, with firms starting some form of RTO, coming in the form of a hybrid setup. In 2022, we see RTO mentions rise even more as companies call back employees, even making office not-optional (for example, X).[9]



(a) Mentions of work from home over time.

(b) Mentions of return to office over time.

Figure 4: Temporal analysis of WFH and RTO options.

To put it definitively, only considering the years 2019 to 2022, we see a significant ($p < 0.05$) and strong negative correlation (r=-0.9976) between WFH and RTO.

3.3.3 Layoffs

For layoffs, we see mentions spike twice – first in 2020 and second in 2022 (Figure 5a). The first can be attributed to the initial global and national economic shocks due to the spread of COVID-19 (Brodeur et al., 2021). This was followed by a short year of massive growth as companies adapted. The second is due to the bubble burst in mid-2022, with companies reporting slower growths due to delayed supply chain disruption effects and announcing layoffs once again. The situation worsened towards the end of 2022 as tech giants like Meta announced their first of many rounds of layoffs.[10] A curious small peak in 2016 can also be observed, which also maps to tech giants layoffs of 2016.[11]

3.3.4 Sentiments

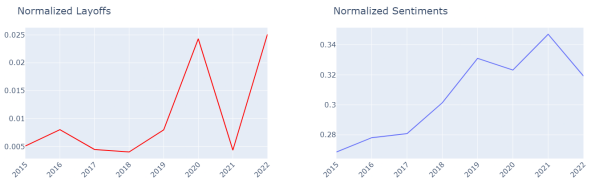
To get a holistic perspective on the opinions of Blind, we annotated the *Blind Posts* for sentiments, using VADER (Hutto and Gilbert, 2014). Each score is extracted from the *compound* field, which combines the sentence's negative and posi-

[8] https://www.computerworld.com/article/3679733/tech-layoffs-in-2022-a-timeline.html

[9] https://fortune.com/2023/03/24/return-to-office-elon-musk-twitter-tesla-layoffs/

[10] https://about.fb.com/news/2022/11/mark-zuckerberg-layoff-message-to-employees/

[11] https://www.cio.com/article/218133/9-bloodiest-tech-giants-layoffs-of-2016.html



(a) Mentions of layoffs over time.

(b) Combined sentiments of the network.

Figure 5: Temporal analysis of layoffs and sentiments.

tive scores. The final value is between -1 (negative sentiment) and +1 (positive sentiment), with 0 as neutral, which is then aggregated and normalized year-wise.

We see the sentiments reflect the broader state of the tech industry in the real world (Figure 5b). 2020 and 2022 are the years of layoffs, as reflected by the sentiment dips. Even as 2020 was the biggest downfall globally, we hypothesize the positive sentiments due to WFH cushioned the fall a little. We see the sentiments at an all-time high in 2021 due to adopting and embracing the new tech. The increased profits and sentiments were consistent across all big corporations.[12] Even the rise of sentiments in 2018 and 2019 can be correlated with the high hiring pace of Meta and Alphabet at the time.[13]

4 Beyond Blind

Our current analysis is entirely restricted to the confines of the Blind platform. However, discussions about the tech industry are spread across other social networks, such as X, Reddit, Quora, and more. The only issue is a plethora of other discussions not related to our interests. As an employee or employer, it would be highly productive to consolidate discussions and opinions from various platforms. This will allow us to get Blind-level insights and extend them to an internet-level analysis. In this section, we propose a novel content classification pipeline that automatically filters out relevant tech content from an ocean of discussions, annotates them with finer classes, and further automatically identifies opinions (Figure 7).

[12] https://www.bloomberg.com/news/articles/2022-03-30/2021-was-best-year-for-u-s-corporation-profits-since-1950

[13] https://www.forbes.com/sites/jackkelly/2023/01/25/tech-layoffs-look-terrible-but-theyre-only-a-pullback-from-years-of-aggressive-hiring/

4.1 Data

4.1.1 Coarse Classification

We first aim to filter tech-related content from any platform by modeling this task as a binary supervised classification task.

Since not all content on Blind is tech-related (there are boards on gaming, entertainment, etc.), we handpick 41 tech-related boards, totaling 664,048 (of 767k) *Blind Posts*. The selected boards are the following – "Work Visa", "Layoffs", "Referrals", "Job Openings", "Work From Home", "Return to Office", "Compensation", "Side Jobs", "Startups", "IPO" and every category with "Industry" or "Career" in its name.

Since the remaining 103k non-tech posts are insufficient for a balanced dataset, we instead utilize the Reddit TL;DR dataset (Völske et al., 2017). Using this dataset has a two-fold advantage. First, the number of posts is high – 3.8M+ posts from various categories like relationship, gaming, advice, movies, politics, etc., allowing us to have enough variety of labeled data points. Second, the content length distribution for Blind and Reddit posts is similar, enabling the model to learn content distribution differences instead of just content length differences. To ensure the validity of the data, we manually go through the largest 100 subreddits and check for any tech-related ones – "sysadmin", "Android", "techsupport", "talesfromtechsupport", and "technology" are thus removed. We sample an equal number of posts (664,048) from the remaining 95 largest subreddits, giving us a total of 1.3M+ labeled data points for this task.

4.1.2 Fine-Grained Classification

Next, we aim to classify any text in one of the ten most popular categories in *Blind Posts*.[14] These boards have more than 10k+ counts each, with only "Tech Industry" and "Software Engineering Career" having 100k+ posts. We sample 10k posts for each category to get the best results.

4.1.3 Opinion Classification

Finally, we aim to extract the opinions of the text automatically by capitalizing on the pro and con fields in *Blind Company Reviews*. This would give us an idea of whether employees view a company

[14] "Tech Industry", "Software Engineering Career", "Work Visa", "Investments & Money", "Housing", "Product Management Career", "Finance Industry", "Referrals", "Data Science & Analytics Career", and, "Compensation"

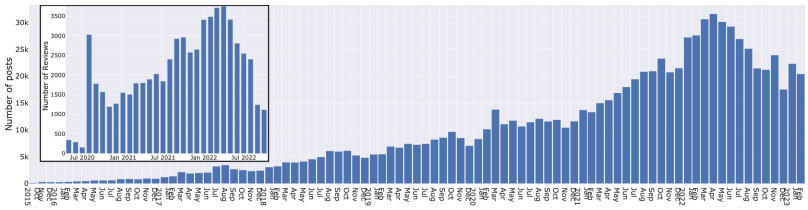


Figure 6: Year-Over-Year activity on Blind using posts (main graph) and reviews (inner graph).

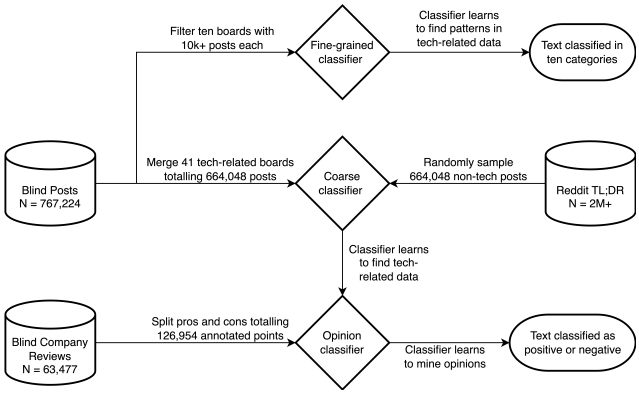


Figure 7: *Blind Posts* and *Blind Company Reviews* classification pipeline.

positively or negatively. Each of the 63,477 reviews is thus split into two, giving us an automatically balanced & labeled dataset of close to 127k data points.

4.2 Methodology & Experiments

4.2.1 Machine Learning Models

We first preprocess the text using NLTK's Tweet-Tokenizer (Bird et al., 2009) to better represent the social-media-esque posts of Blind. This is followed by a tf-idf vectorizer to embed count-based significance to each word and send it to machine learning (ML) models. We establish baselines using models that have proven effective for text classification tasks, namely, Logistic Regression, Linear-Support Vector Classifier (Linear-SVC), and Multinomial Naive Bayes (Multinomial NB) (Aggarwal and Zhai, 2012).

4.2.2 Transformer Models

Transformers have surpassed ML models in text classification tasks (Vaswani et al., 2017). By generating contextual semantic embeddings instead of static syntactic embeddings, rigorous model pre-training, and leveraging attention ensures they have a nuanced understanding of the language.

Specifically, we train the base uncased models of the following models. *BERT* (Bidirectional Encoder Representations from Transformers) is a pre-trained transformer-based model that captures contextual embeddings from forward and backward directions, improving natural language understanding (NLU) tasks (Devlin et al., 2018). *DistilBERT* is a smaller, distilled version of BERT, applying knowledge distillation to achieve comparable performance with fewer parameters (Sanh et al., 2019). *RoBERTa* (A Robustly Optimized BERT Pretraining Approach) is an extension of BERT, using larger batch sizes and more data, resulting in improved performance and generalization across various NLU tasks (Liu et al., 2019).

We fine-tune the transformers on the datasets for two epochs using the AdamW optimizer, varying the learning rate between 2e-5 and 5e-5 for the best results. We train all the above ML and transformer models using 5-fold cross-validation.

4.3 Results & Discussion

All the results are summarized in Table 2.

4.3.1 Coarse Classification

For the coarse classification task, we achieved a stellar accuracy of 99.25% with the linear support vector classifier model. The high accuracy is especially significant given the dataset is balanced. Post-hoc analysis of the model shows that it has learned named entities (NEs) and popular keywords. For example, "After spending almost 5 years at Facebook ..." is correctly classified as

Classification Results	**Coarse Content Classification**		**Fine-Grained Content Classification**		**Company Review Classification**	
Model ↓ / Metric →	F1	Accuracy	F1	Accuracy	F1	Accuracy
Logistic Regression	0.9796	0.9796	0.7232	0.7224	0.9438	0.9434
Linear-SVC	**0.9925**	**0.9925**	0.7229	0.7161	0.9427	0.9427
Multinomial NB	0.9797	0.9797	0.6962	0.6852	0.9220	0.9208
DistilBERT	Accuracy already reached >99%.		0.7674	0.7734	0.9775	0.9777
BERT			0.7772	0.7822	0.9783	0.9786
RoBERTa			**0.7780**	**0.7841**	**0.9828**	**0.9829**

Table 2: Results for all the classification tasks.

tech, and switching "Facebook" to "CSGO" (a popular game) shows non-tech. Similarly, for "I was removed from the company", we see the model correctly classifies this as tech. Changing "company" to "community" again flips the category correctly. Since the model is purely syntactic, these results show that the model has learned the distinctive distributions of keywords and NEs in the tech and non-tech parts of the dataset.

4.3.2 Fine-Grained Classification

For the fine-grained classifier, we see that the accuracies of ML models are lower. This can be attributed to two reasons. First, the number of categories is higher (ten instead of two), and second, the model can no longer pick up on the broader themes of the tech industry – since all the data lies under that distribution. Here, we see significant jumps in accuracy from traditional ML models to state-of-the-art transformer models, reaching 78.41% accuracy with RoBERTa.

4.3.3 Opinion Classification

We see a similar jump in accuracy for the opinion classification task, reaching a 98.29% score with the RoBERTa model. It is worth noting that even without the transformers, ML models show a high accuracy of 94.34%. This might be due to the different frequencies of adjectives in pros and cons. We would find a pro is more likely to contain "good", "great", "decent", etc., and a con is more likely to contain "bad" (management/compensation), "toxic" (culture), and "slow" (growth)

All ML models have the upper limit of understanding the language's basic syntax since their embeddings are generated using count-based methods (tf-idf). With the attention mechanism and rigorous pretraining, transformers gain the ability to touch the semantics of language – thus showing us accuracy jumps.

4.4 Limitations

4.4.1 Linguistic Bottlenecks

When it comes to opinions, failure to recognize humor devices of sarcasm and irony remains a challenge (Gregory et al., 2020). For example, "You never work a day in your life if your hobby is your work" is an ambiguous review and can be interpreted in any way. Similarly, "Experience is whatever you make of it, but if you're driven, you can transition into other departments." is ambiguous.

Other trickier instances fail to get recognized by even the transformer models. Consider the following pro review: "Work and Life balance because managements are not working. Just having inefficient meetings".[15]

4.4.2 Dataset Nuances

Consider the text "My parents are getting me married ...". One would expect this to be present in a relationship advice subreddit and thus should be labeled as non-tech. Yet, the coarse classifier labels this as a tech-related post. Partly, it might be due to users incorrectly selecting the boards. It might also be due to the growing Indian population on the platform, where marriage is considered a primary landmark of life.[16]

4.4.3 Biased Population

It is well known that selecting any platform for analysis would introduce its biases (Ferrer et al., 2021; Cihon and Yasseri, 2016). For example, Tata Consultancy Services (TCS), the biggest tech industry employer in India, employs 528k+ people.

[15] It is *because* the management is not working and having inefficient meetings, that the user has balance between work and life.

[16] https://www.bbc.com/news/world-asia-india-59530706

Google, a multi-national company, has around a fifth (140k+) of employees for comparison. Yet, the number of *Blind Company Reviews* for Google is 7.8k, much higher than that of TCS, with only 573 reviews. It is critical to note that Blind represents the privileged sector of the tech industry.

5 Related Work

5.1 Online Social Networks

Plenty of work has been done to understand OSNs. Large-scale studies on Facebook (Wilson et al., 2012), X (Kwak et al., 2010), Reddit (Proferes et al., 2021), and more have been conducted to understand the science behind how networks work and interact. At the same time, there is a growing privacy concern about social media user profiling. This has led to a unique social network, where the network is formed purely by the content of interactions, and the identity is anonymized.

5.2 Blind

Past work on Blind is extremely limited and focuses prominently on the anonymity aspect. Perceptions and uses of anonymity in IT organizations are explored using a poll for Microsoft employees (Kim and Scott, 2018). The authors extend their work further and find communication qualities and freedom of speech at work play a significant role in the work environment (Kim and Scott, 2019; Kim and Leach, 2020). A more recent paper attempts to find evidence of the usefulness of earnings announcements to job-seekers (Choi et al., 2023). Therefore, our work on the last seven years of Blind data becomes the platform's first large-scale empirical text-based analysis.

5.3 Text Classification

There is a heap of work on text classification, from tasks like automatic movie sentiment analysis (Baid et al., 2017), humor detection (Chaudhary et al., 2021), fraud detection (Singh et al., 2022), and more. Earlier works on text classification focus on machine learning approaches (Ikonomakis et al., 2005). We include those as baselines and improve using transformer-based architectures like BERT (Devlin et al., 2018), and its variants DistilBERT (Sanh et al., 2019) and RoBERTa (Liu et al., 2019), as explored in the papers (González-Carvajal and Garrido-Merchán, 2020; Minaee et al., 2021).

6 Conclusion

Blind is an anonymous social media platform for professionals, serving the need for the missing negative side of the story in career discussions. In this paper, for the first time, we acquire and conduct large-scale analyses on two novel datasets: 63,477 *Blind Company Reviews* and 767,224 *Blind Posts*, containing seven years of anonymized industry data. User-given company reviews show an abundance of opposing opinions, confirming the unique dual discourse of the platform. We see surprising correlations between metrics across companies, as users talk about work-life balance the most, but we find culture and management as more vital reasons affecting ratings. The views on a post follow the expected Gibrat's rule of proportionate growth. Uniquely, we see an inversion in the number of likes versus comments received, a symptom of anonymity fostering discussions. Exploiting Blind's bias towards the tech industry, we conduct a temporal analysis and find mappings from global trends like COVID-19, work-from-home, return-to-office, and layoffs.

Next, we propose a novel content classification pipeline leveraging the Blind datasets to go beyond. We first filter out tech-related content from an ocean of discussions from any social media with a 99.25% accuracy. Then, we automatically annotate the tech data with ten finer classes, achieving an accuracy of 78.41%, showing a deeper understanding of the text. Finally, we mine the opinions of users to provide an aggregated birds-eye view of the sentiments in the industry, with a 98.29% accuracy. These high-accuracy results demonstrate the high practicality of our novel pipeline.

7 Future Work

With the datasets, a large language model could be trained that offers automatic advice based on the constructive and honest opinions of the Blind community.

Efforts in named entity recognition (NER) can also be taken forward as *Blind Posts* contain an ample frequency of companies, compensation, dates, and other named entities.

References

Charu C Aggarwal and ChengXiang Zhai. 2012. A survey of text classification algorithms. *Mining text*

data, pages 163–222.

Thomas Aichner, Matthias Grünfelder, Oswin Maurer, and Deni Jegeni. 2021. Twenty-five years of social media: a review of social media applications and definitions from 1994 to 2019. *Cyberpsychology, behavior, and social networking*, 24(4):215–222.

Palak Baid, Apoorva Gupta, and Neelam Chaplot. 2017. Sentiment analysis of movie reviews using machine learning techniques. *International Journal of Computer Applications*, 179(7):45–49.

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The pushshift reddit dataset. In *Proceedings of the international AAAI conference on web and social media*, volume 14, pages 830–839.

Alexander Bick, Adam Blandin, Karel Mertens, et al. 2020. Work from home after the covid-19 outbreak.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Abel Brodeur, David Gray, Anik Islam, and Suraiya Bhuiyan. 2021. A literature review of the economics of covid-19. *Journal of economic surveys*, 35(4):1007–1044.

Tanishq Chaudhary, Mayank Goel, and Radhika Mamidi. 2021. Towards conversational humor analysis and design. *arXiv preprint arXiv:2103.00536*.

Bong-Geun Choi, Jung Ho Choi, and Sara Malik. 2023. Not just for investors: The role of earnings announcements in guiding job seekers. *Journal of Accounting and Economics*, page 101588.

Peter Cihon and Taha Yasseri. 2016. A biased review of biases in twitter studies on political collective action. *Frontiers in Physics*, page 34.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Xavier Ferrer, Tom van Nuenen, Jose M Such, and Natalia Criado. 2021. Discovering and categorising language biases in reddit. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, pages 140–151.

Santiago González-Carvajal and Eduardo C Garrido-Merchán. 2020. Comparing bert against traditional machine learning text classification. *arXiv preprint arXiv:2005.13012*.

Hunter Gregory, Steven Li, Pouya Mohammadi, Natalie Tarn, Rachel Draelos, and Cynthia Rudin. 2020. A transformer approach to contextual sarcasm detection in twitter. In *Proceedings of the second workshop on figurative language processing*, pages 270–275.

Clayton Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, volume 8, pages 216–225.

M Ikonomakis, Sotiris Kotsiantis, and Vasilis Tampakas. 2005. Text classification using machine learning techniques. *WSEAS transactions on computers*, 4(8):966–974.

David E Kanouse, L Hanson, Edward E Jones, H Kelley, R Nisbett, S Valins, and B Weiner. 1972. Attribution: Perceiving the causes of behavior. *Morristown, NJ: General Learning*, pages 47–62.

Heewon Kim and Rebecca Leach. 2020. The role of digitally-enabled employee voice in fostering positive change and affective commitment in centralized organizations. *Communication Monographs*, 87(4):425–444.

Heewon Kim and Craig Scott. 2019. Change communication and the use of anonymous social media at work: Implications for employee engagement. *Corporate Communications: An International Journal*.

Heewon Kim and Craig R Scott. 2018. Going anonymous: Uses and perceptions of anonymous social media in an it organization. In *Proceedings of the 9th International Conference on Social Media and Society*, pages 335–339.

Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. 2010. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, pages 591–600.

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*.

Edwin Mansfield. 1962. Entry, gibrat's law, innovation, and the growth of firms. *The American economic review*, 52(5):1023–1051.

Shervin Minaee, Nal Kalchbrenner, Erik Cambria, Narjes Nikzad, Meysam Chenaghlu, and Jianfeng Gao. 2021. Deep learning–based text classification: a comprehensive review. *ACM computing surveys (CSUR)*, 54(3):1–40.

Nicholas Proferes, Naiyan Jones, Sarah Gilbert, Casey Fiesler, and Michael Zimmer. 2021. Studying reddit: A systematic overview of disciplines, approaches, methods, and ethics. *Social Media+ Society*, 7(2):20563051211019004.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Ashwin Singh, Arvindh Arun, Pulak Malhotra, Pooja Desur, Ayushi Jain, Duen Horng Chau, and Ponnurangam Kumaraguru. 2022. Erasing labor with labor: Dark patterns and lockstep behaviors on google play. In *Proceedings of the 33rd ACM Conference on Hypertext and Social Media*, pages 186–191.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Michael Völske, Martin Potthast, Shahbaz Syed, and Benno Stein. 2017. Tl; dr: Mining reddit to learn automatic summarization. In *Proceedings of the Workshop on New Frontiers in Summarization*, pages 59–63.

Robert E Wilson, Samuel D Gosling, and Lindsay T Graham. 2012. A review of facebook research in the social sciences. *Perspectives on psychological science*, 7(3):203–220.