Mining Insights from Large-scale Corpora Using Fine-tuned Language Models

Shriphani Palakodety[12] and **Ashiqur R. KhudaBukhsh**[13] and **Jaime G. Carbonell**[4]

Abstract. Mining insights from large volume of social media texts with minimal supervision is a highly challenging Natural Language Processing (NLP) task. While Language Models' (LMs) efficacy in several downstream tasks is well-studied, assessing their applicability in answering relational questions, tracking perception or mining deeper insights is under-explored. Few recent lines of work have scratched the surface by studying pre-trained LMs' (e.g., `BERT`) capability in answering relational questions through "fill-in-the-blank" cloze statements (e.g., `[Dante was born in MASK]`). `BERT` predicts the MASK-ed word with a list of words ranked by probability (in this case, `BERT` successfully predicts Florence with the highest probability). In this paper, we conduct a feasibility study of fine-tuned LMs with a different focus on tracking polls, tracking community perception and mining deeper insights typically obtained through costly surveys. Our main focus is on a substantial corpus of video comments extracted from YouTube videos (6,182,868 comments on 130,067 videos by 1,518,077 users) posted within 100 days prior to the 2019 Indian General Election. Using fill-in-the-blank cloze statements against a recent high-performance language modeling algorithm, `BERT`, we present a novel application of this family of tools that is able to (1) aggregate political sentiment (2) reveal community perception and (3) track evolving national priorities and issues of interest.

1 INTRODUCTION

Pre-trained Language Models (LMs), such as `BERT` [14], ELMo [26], XLNet [38] etc. have received widespread attention in recent NLP literature. While Language Models' (LMs) efficacy in several downstream tasks is well-studied, assessing their applicability in answering relational questions is largely under-explored. Recent lines of work have begun to scratch the surface with analyzing LMs' capability in answering relational questions presented as "fill-in-the-blank" cloze statements. Competing views about their effectiveness as Knowledge Bases (KBs) have been published [21, 27].

While the jury is still out on how effective LMs are as Knowledge Bases in their current form, in this paper, we explore a related research question: *is it possible to track community perception, aggregate opinions and compare popularity of political parties and candidates using LMs?*

In this paper, we introduce a YouTube comment corpus relevant to the Indian General Election (6,182,868 comments on 130,067 videos by 1,518,077 users). `BERT` is pre-trained on a book corpus and Wikipedia, i.e., on well-formed texts by contributors proficient in English covering a broad range of topics [14]. In contrast, our election corpus consists of short texts with grammar and spelling disfluencies and has a topical focus on the general election. In a series of experiments using cloze statements, we demonstrate that, in its current form, fine-tuned `BERT` can shed interesting insights into three previously unexplored tasks in the context of knowledge-mining using LMs: (1) community perception analysis, (2) comparative analysis of popularity of candidates or political parties, and (3) mining deeper insights about national priorities. We side-step known issues of handling negative cloze statements [21] with corpus modification and construct interesting adversarial scenarios to guide our intuitions better.

In social science, major studies often rely on extensive surveys. Typically, such surveys are few-and-far-between as conducting them on a regular basis involves significant resources. Also, aggregating opinions at multiple spatiotemporal granularities is a non-trivial challenge. Further, language modeling and querying allows us to sidestep issues of knowledge schema engineering and complex modeling to integrate this schema. In this work, we investigate the possibility of using `BERT` to complement traditional surveys. Note that, our findings are not limited to the current success reported in this paper. We are rather making a more general claim that going forward, LMs can provide a compelling solution for performing fast-turnaround analysis while requiring minimal supervision.

Contributions: Our contributions are the following[5].

1. ***Social:*** To the best of our knowledge, we report the first large-scale social media analysis of community perception focused on two major religions in India. Religion has remained a contentious issue both during the pre-independence (1947) [33] and post-independence era [16, 29] in India. Our analysis provides a starting point for further research and journalism in this direction [8, 7, 4].
2. ***High-performance Language Modeling for insight mining:*** We investigate the capabilities of a high-performance language modeling tool, `BERT`, and present an exploratory study on the model's capability to reveal a variety of insights that align well with actual observations, outcomes, and surveys.
3. ***Side-stepping known issues, analysis of retention:*** We propose a corpus modification solution to side-step a recently-reported issue with handling negated cloze statements. We report a new analysis outlining how much knowledge a fine-tuned `BERT` *retains* and provide interesting insights.

[1] Ashiqur R. KhudaBukhsh and Shriphani Palakodety are equal contribution first authors. Ashiqur R. KhudaBukhsh is the corresponding author.
[2] Onai, USA, email: spalakod@onai.com
[3] Carnegie Mellon University, USA, email: akhudabu@cs.cmu.edu
[4] Carnegie Mellon University, USA, email:jgc@cs.cmu.edu

[5] Resources and additional details are available at: https://www.cs.cmu.edu/~akhudabu/BERT2019IndianElection.html

2 DATA: YOUTUBE COMMENTS

YouTube channels: We considered YouTube channels for 11 highly popular national news outlets and 3 highly popular national newspapers' official YouTube channels. Of the 29 states in India, we restricted our focus on 12 states (listed in Table 1) that contribute 20 or more seats in the lower house of parliament. Our analysis encompasses a large fraction of the political voice in India since these 12 states account for 423 seats out of 543 seats, i.e., 77.9% of the total seats in the parliament. In terms of vote share, of the overall 613,133,300 votes cast in 2019 election, 534,378,886 (87.16%) votes were cast from these 12 states [3]. For each of the states, we identified two highly popular YouTube news channels. Overall, this implies 38 YouTube channels (24 regional, 14 national). The average subscriber count of these channels is 3,338,628 (average subscriber count for national YouTube channels: 5,840,950 ; average subscriber count for regional YouTube channels: 1,878,941).

Andhra Pradesh (25), Bihar (40), Gujarat (26), Karnataka (28), Kerala (20), Madhya Pradesh (29), Maharashtra (48). Odisha (21), Rajasthan (25), Tamil Nadu (39), Uttar Pradesh (80), West Bengal (42)

Table 1: States with seat counts in brackets.

Period of interest: We considered a 100 day period starting from Feb 12th to May 22nd, 2019. This spans a 100-day window preceding the announcement of results.

Characterization of the videos: Our video data set, $\mathcal{V}$, consists of 130,067 videos uploaded in the 38 YouTube channels during our period of interest (46,055 videos from national channels, 84,012 videos from regional channels). It is not feasible to manually label all videos as relevant (i.e., talking about some aspects concerning the Indian election) or irrelevant (i.e., talking about unrelated topics like entertainment, sports etc.). We randomly selected 100 videos and manually annotated them with the following labels: politics, entertainment, sports, weather, crime, finance and others. Note that, we do not intend these categories to be formal or exhaustive, but rather to be illustrative of the types of news videos that were uploaded during our period of interest and provide a rough estimate of the relative distribution of political news in our video data set. As seen in Table 2, a substantial chunk of the news videos were on politics mainly covering the election updates, debates among party spokespersons, evaluation of campaign promises and foreign policy discussions.

Categories	# of videos
Politics	72
Weather	3
Entertainment	3
Crime report	2
Finance	1
Sports	1
Other	18

Table 2: Characterization of sampled YouTube channels.

Comments data set: Using the publicly available YouTube API, we crawled the comments posted on videos in $\mathcal{V}$. Overall, we obtained 6,182,868 comments (4,198,599 comments from national channels, 1,984,269 comments from regional channels). One major impediment to analyzing social media responses generated in the Indian subcontinent is its linguistic diversity. We used a recently-proposed, high-accuracy polyglot embedding based language identification technique (first proposed in [24] and successfully replicated in a different multilingual corpus [25]) to separate the English corpus. Overall, we obtained 1,940,757 English comments (denoted as $\mathcal{C}_{all}$) with the following breakdown: 1,512,009 comments from the 14 national YouTube news channels (denoted as $\mathcal{C}_{national}$), 428,748 comments from the 24 regional YouTube channels.

Preprocessing: We follow the standard preprocessing steps recommended for the `BERT` [14] language model for our fine-tuning tasks. For our task we use the uncased English model with the following parameter details: 12 transformer layers, hidden state length of 768, 12 attention heads, 110M overall parameters[6]. Note that these parameters are recommended by the authors of `BERT` and our analysis shows that they work well for our task as well. The base `BERT` vocabulary is supplemented by 900 most frequent tokens from the English subset of our corpus. Finally, the pre-trained model is fine-tuned on the target corpus in question using the training hyperparameters are presented below.

- Batch size: 16
- Maximum sequence length: 128
- Maximum Predictions Per Sequence: 20
- Fine-tuning steps: 20,000
- Warmup steps: 10
- Learning rate: 2e-5

2.1 A Challenging Data Set

Similar to most data sets of short social media texts generated in a linguistically diverse region, our data set exhibits a considerable presence of out-of-vocabulary (OOV) words, code-mixing, and grammar and spelling disfluencies. In addition to these challenges, given that a vast majority of the content contributors does not speak English as their first language, we noticed a substantial incidence of phonetic spelling errors (e.g., [`human beings are important not vehicles are` **`bloody pupil`**] originally intended to express **bloody people**); 32.67% of times, the word `liar` was misspelled as `lier`. In all, considering terms occurring 5 or more times in the corpus, the OOV rate against the $\texttt{BERT}_{base}$ was 75.08%.

To summarize, our data set (i) captures a considerable fraction of political voice of India (ii) is obtained from videos predominantly discussing election (see, Table 2) and (iii) is markedly different from the documents used to train the original $\texttt{BERT}_{base}$.

3 RELATED WORK

Election analysis: Social media analysis of a variety of elections across several countries has been widely studied (e.g., US [35, 15, 23], UK [10], India [19, 30, 22], Netherlands [28], Pakistan, South Korea [31] etc.); presenting an exhaustive analysis is beyond the scope of this paper. Special referendum elections like Brexit [11] and the Greek Referendum [34] have also received attention from the Information Retrieval (IR) community. Three major directions distinguish our work from prior literature: (1) our focus on YouTube comments, a rather under-explored data resource instead of twitter vast majority of previously published work focused on Twitter (2) `BERT` predictions instead of previously explored signals like lexicon-based sentiment analysis, tweet volume, tweet mentions etc. to correlate with election outcome (3) beyond typical comparative analysis of popularity measures of candidates and parties, a broader scope to track evolving national priorities, and community perception.

Sentiment-mining using `BERT`: Unrelated to the task of political sentiment-mining, in terms of sentiment analysis using `BERT`, the

[6] https://github.com/google-research/bert/

closest work to our contribution is a targeted aspect-based sentiment analysis (TABSA) task presented in [32]. Our work is different along the following lines. First, our focus is on political text-mining as opposed to the TABSA task. Second, our data set is substantially more challenging than the Sentihood data set used in [32] indicating BERT's robustness to noisy social media text generated in a part of the globe where the majority of the content contributors are non-native speakers of English. Finally, and most importantly, beyond sentiment-mining of political actors, we mine deeper insights from the corpus such as evolving national priorities and track community perception.

Language models as Knowledge Bases (KBs): Recent attempts to answer relational questions using LMs have recieved moderate success by casting the relational questions as "fill-in-the-blank" cloze statements (e.g., `[Gordon Scholes is a member of the MASK]` political party. - expected answer Labor) [27]. However, further probing of these models has uncovered limitations in their handling of negated cloze statements [21]. For instance, these models often tend to provide near-identical answers to negated queries e.g., when `[Birds cannot MASK]` and `[Birds can MASK]` are used as cloze statements, the answer fly is predicted with high probability in both cases. Our work is different in the following ways. First, unlike [27, 21], we work with fine-tuned BERT operating on a challenging corpus of social media texts produced mostly by non-native speakers of English. Second, instead of answering relational queries, we focus on tracking community perception, mining national priorities and comparing relative popularity of political entities. Third, we provide a comparative analysis demonstrating the extent to which a fine-tuned BERT language model forgets the base knowledge contained in the original pre-trained model (i.e. retention) - a key aspect to consider if LMs have to replace KBs. Finally, we propose a solution to sidestep the issue of negated queries by removing documents (comments) containing valence shifters.

4 BACKGROUND

BERT: BERT [14] is a recent high-performance bi-directional transformer language model. The transformer architecture is a recent deep neural model for sequence-to-sequence prediction tasks. A sequence-to-sequence task involves accepting a sequence as input and producing a sequence as output. Models for tackling these problems typically contain an encoder that operates on the input and constructs a representation, and a decoder that operates on the representation (and also the input in some cases) and produces the desired output. Both the encoder and decoder in the transformer model use the Multi-Head attention mechanism to *attend* to different input positions.

Large scale language models, trained on large corpora, have recently produced strong results in text generation, and strong downstream performance for tasks like text-classification. BERT itself has produced significant performance-gains in a slew of NLP tasks [14]. BERT uses a transformer model [36] with a masked-word prediction objective and a next sentence prediction auxiliary training objective.

Recent work has explored the knowledge present in these (not fine-tuned) large scale language models using cloze sentences [27, 21]. As shown in Figure 1, we evaluate the fine-tuning paradigm on an Indian election corpus. In Section 6, we provide an analysis of the acquisition and retention properties of fine-tuned BERT.

Indian election: India follows a multi-party parliamentary system. The general election allows the voter-base to elect the 543 members of the lower house of parliament - The Lok Sabha. The winning party or a coalition of parties then nominate one of the members to serve as

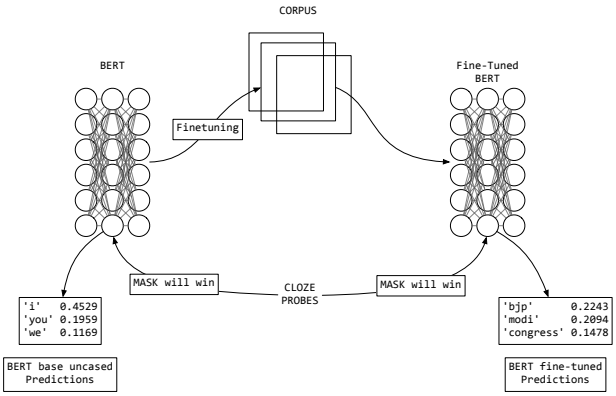


Figure 1: System diagram.

the Prime Minister. The 2019 election was conducted over a period starting 11^{th} of April and ending on the 19^{th} of May in 7 phases. The votes were counted and the results were announced on the 23^{rd} of May. The ruling party (BJP) won an outright majority and Narendra Modi - the incumbent prime minister was nominated for a second term. In our work, we focus on two major political parties: Indian National Congress (popularly, referred to as Congress) and Bharatiya Janata Party (popularly, referred to as BJP) and the two projected prime-ministerial candidates: Narendra Modi, and Rahul Gandhi.

5 RESULTS AND ANALYSIS

5.1 Sanity check on fine-tuning

We denote finetuned BERT on $\mathcal{C}_{national}$ and $\mathcal{C}_{all}$ as $\text{BERT}_{national}$ and BERT_{all}, respectively. Since BERT_{base} is pre-trained on a book corpus and Wikipedia data set, it is possible that without any fine-tuning, it may reflect information relevant to India. For example, when the input sentence is `[MASK is a major Indian city]`, BERT's top three predictions are Chennai, Delhi and Mumbai with probabilities 0.15, 0.12 and 0.12, respectively. However, the results could be slightly dated.

For instance, Table 3 presents the top three completions ranked by probability on the following cloze statements:

- `[MASK Gandhi]` (denoted as $cloze_1$)
- `[Narendra MASK]` (denoted as $cloze_2$)

BERT_{base} on $cloze_1$ included two deceased former Indian politicians belonging to the Gandhi family: Indira Gandhi (former prime minister of India), Sanjay Gandhi (son of Indira Gandhi and also a politician). In contrast, both fine-tuned $\text{BERT}_{national}$ and BERT_{all} predicted the currently active politicians from the same family. Moreover, on $cloze_2$, BERT_{base} failed to suggest Modi, the most obvious completion in contemporary Indian politics. Our test indicates that on simple cloze statements, fine-tuned BERT outputs results consistent with the corpus.

Probe	BERT_{base}	$\text{BERT}_{national}$	BERT_{all}
$cloze_1$	Indira (0.82), Sonia (0.04), Sanjay (0.01)	Rahul (0.58), Fake (0.08), Priyanka (0.05)	Rahul (0.6), Priyanka (0.04), Sonia (0.03
$cloze_2$	Kumar (0.16), Sharma (0.14), Singh (0.07)	Modi (0.77), Modiji (0.02), sir (0.01)	Modi (0.70), Modiji (0.02), Rahul (0.01)

Table 3: Predicted completions with probabilities in parentheses.

5.2 Community perception tracking

Research question: *Can we use fine-tuned LMs to track community perception?* A trend of increasing polarization in the Indian political scene along religious lines has been reported recently [8, 7, 4]. Analysis of religious polarization in our corpus (along the lines of political polarization in [13]) would require a reliable estimate of religious affiliation. Hence, instead we focused on the tracking perception of the two prominent religions in India. We conduct several modifications to our corpora to eliminate possibilities of inaccurate characterization and employ different techniques to analyze this research question of considerable social value.

We first construct a simple test to ascertain that fine-tuned BERT reflects discussions around religion in the corpus. In response to the cloze sentence [My religion is MASK], the top two BERT_{base} predictions are Christian and Catholic while the fine-tuned BERT_{all} predicts Islam and Hindu - in line with expectations. We next construct two cloze statements: [Hindus are MASK] (denoted as $\mathcal{S}_1$) and [Muslims are MASK] (denoted as $\mathcal{S}_2$), and query $\text{BERT}_{national}$, BERT_{all} and BERT_{base} to estimate the perception of these two religions. Among 4,381,623 unique bigrams, in terms of frequency, [Hindus are] and [Muslims are] rank 755^{th} and 699^{th}, respectively.

Table 6 lists the top three completions suggested by different BERT models. Our findings highlight the following points. First, fine-tuning substantially altered the predictions; with BERT_{base}, both $\mathcal{S}_1$ and $\mathcal{S}_2$ were completed with predominantly neutral terms. However, a marked shift in the nature of completions was observed with models trained on the election corpus; the top-predicted words for both $\mathcal{S}_1$ and $\mathcal{S}_2$ were largely negative. Second, the negativity is not merely one-sided – i.e., it is not the case that only one community is painted with negative words while the other is hardly at the receiving end. Rather, both communities received a comparable share of negative completions and almost mirrored each other hinting at a possible polarized political landscape based on religious identities.

Research question: *Is this analysis affected by BERT's inability to account for negation?* It is possible that BERT's predictions are influenced by a prevalence of phrases containing negation (e.g., Hindus are not fools, Muslims are not terrorists). We queried the models obtained by finetuning on the corpora after removing any comment ($\sim$20% of the corpus) containing one or more valence shifters (listed in Table 4). We found the orders of results were unchanged.

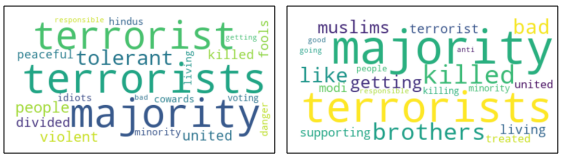
not, can't, won't, don't, shouldn't, mustn't, should not, must not, do not, cannot, will not, would not, wouldn't, isn't, is not, dare not, have not, might not, may not, need not, ought not, shall not

Table 4: List of valence shifters we considered.

Research question: *Is it possible that the analysis is affected by association between the two religious entities?* In a later result, we found that BERT developed an intuition that Modi is related to BJP. We were curious to know if the mirroring predictions of the two communities' perception is influenced by association, i.e., BERT figures out that Hindus and Muslims are related and then one community's perception is reflected on the other. To eliminate this possibility, we further modified the corpora without any negation: (i) holding out all comments containing at least one high frequency term related to Hinduism (listed in Table 7) ($\mathcal{C}_{all}^{Hindu}$) (ii) holding out all comments containing at least one high frequency term related to Islam (Table 7 $\mathcal{C}_{all}^{Islam}$. As shown in Table 5, our results are consistent with Table 6.

$\text{BERT}_{national}^{Muslim}$	$\text{BERT}_{national}^{Hindu}$	$\text{BERT}_{all}^{Muslim}$	$\text{BERT}_{all}^{Hindu}$
$\mathcal{S}_1$	$\mathcal{S}_2$	$\mathcal{S}_1$	$\mathcal{S}_2$
fools (0.09)	fools (0.08)	fools (0.10)	fools (0.07)
terrorists (0.06)	terrorists (0.06)	terrorists (0.06)	terrorists (0.06)
idiots (0.03)	stupid (0.02)	fool (0.05)	stupid (0.02)

Table 5: BERT completion results for [Hindus are MASK] (denoted as $\mathcal{S}_1$) and [Muslims are MASK] (denoted as $\mathcal{S}_2$).



(a) Hindus are (b) Muslims are

Figure 2: A word cloud visualization of [Hindus are] and [Muslims are].

5.2.1 Validation

Using a template-based word-cloud tool (results presented in Figure 2), a semantic lexicon induction tool, and manual inspection, (discussed later) we corroborate BERT's finding that the community perception of both religions was largely negative.

Sentiment analysis using SENTPROP [17]: Lexicon-based sentiment analysis is a well-established method for computing sentiment scores of documents [23]. In this scheme, tokens are assigned scores and individual documents' (comments in our case) scores are obtained by combining the constituent token scores (usually by simple addition). For effective sentiment analysis, obtaining a domain-specific lexicon is crucial [37]. We induced a custom lexicon from our own corpus using word embeddings trained with [18] and a lexicon inducing algorithm (SENTPROP) from [17]. We used the same set of seed words presented in [17] and our test for a positive or negative comment simply adds the individual token scores and if the cumulative comment score is greater than 3 (or less than -3), the comment is considered positive (or negative).

We identified four high-frequency religious tokens for both religions (Hindu, Hindus, Hinduism, Hindutva, Muslim, Muslims, Islam and Islamic) and list their scores in Table 7. We noticed that the scores for all these terms were negative and comparable across both religions. For $\mathcal{C}_{national}$, 71.2% of the tokens were more positive than any of the religious tokens we considered. For $\mathcal{C}_{all}$, 88.2% of the tokens were more positive than any of the religious tokens we considered.

We next divide both $\mathcal{C}_{all}$ and $\mathcal{C}_{national}$ into two disjoint subsets. A *religious* subset containing comments with at least one of the eight religious tokens we considered and *religious*^{c} its complement. In Table 8, we present the percentage of positive and negative comments in the *religious* and *religious*^{c} subsets. We found that compared to the *religious*^{c} subset, the relative increase in fraction of negative comments was more than the relative increase in fraction of positive comments in the corresponding *religious* subset. We performed the same analysis at a finer granularity of individual months. Our finding was consistent; religious discussion attracted more negativity than positivity. We cannot come to a strong conclusion based on our findings, however, in addition to automated analysis, we sampled 100 comments from both *religious* and *religious*^{c} and our manual inspection aligns with our current findings.

BERT$_{base}$	BERT$_{base}$	BERT$_{national}$	BERT$_{national}$	BERT$_{all}$	BERT$_{all}$
$\mathcal{S}_1$	$\mathcal{S}_2$	$\mathcal{S}_1$	$\mathcal{S}_2$	$\mathcal{S}_1$	$\mathcal{S}_2$
here (0.09)	Christians (0.11)	fools (0.15)	fools (0.11)	fools (0.13)	fools (0.09)
minority (0.06)	excluded (0.04)	terrorists (0.07)	terrorists (0.07)	terrorists (0.05)	terrorists (0.06)
Christians (0.06)	Muslim (0.04)	fool (0.02)	fool (0.02)	idiots (0.03)	terrorist (0.03)

Table 6: BERT completion results for [Hindus are MASK] (denoted as $\mathcal{S}_1$) and [Muslims are MASK] (denoted as $\mathcal{S}_2$). Among 4,381,623 unique bigrams, in terms of frequency, [Hindus are] and [Muslims are] rank 755^{th} and 699^{th}, respectively.

Token	$\mathcal{C}_{all}$	$\mathcal{C}_{national}$
Hindu	-0.78	-0.58
Hindus	-0.79	-0.57
Hindutva	-0.71	-0.51
Hinduism	-0.99	-0.59
Muslim	-0.72	-0.49
Muslims	-0.80	-0.49
Islam	-0.91	-0.58
Islamic	-0.68	-0.59

Table 7: Sentiment of religious tokens.

	religious subset	*religious*^{c} subset
$\mathcal{C}_{national}$	pos = 35.16% neg = 18.55%	pos = 27.87% neg = 6.24%
$\mathcal{C}_{all}$	pos = 33.64% neg = 18.55%	pos = 18.47% neg = 4.13%

Table 8: Sentiment analysis by partitioning the corpus into subsets containing religious tokens and its complement.

Presence of hate words: In our third and final analysis, we focus on presence of hate tokens around the religious tokens. In the *religious* subset of $\mathcal{C}_{all}$, the aforementioned religious tokens appeared 151,919 times in 95,638 comments (3.94% of the entire corpus). For every such instance of a religious token in a comment, we considered a left and right context of two words (i.e. a total of four surrounding words) around the religious token and computed the fraction of total instances that contained a hate word or a slur. For hate words, we considered a combination of two previously-published lexicons [9, 20] of derogatory terms used in code-switched English (365 unique slurs). We found that at least one slur was present in 7.22% of the contexts containing a religious token.

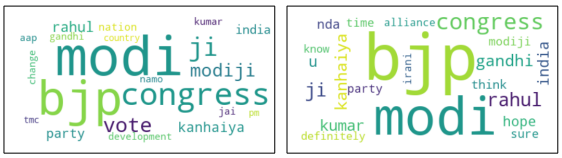
It may very well be the case that terms in these hate lexicons are a common occurrence in Indian online discussions independent of the subject (religious or otherwise). In order to verify if that is the case, we randomly sampled equal number of 4-grams (sequence of 4 consecutive tokens) from the *religious*^{c} subset of $\mathcal{C}_{all}$ and found that the fraction of contexts containing a hate word ($4.39 \pm 0.07\%$) was less indicating that when religion is discussed the presence of hateful terms increases.

5.3 Comparing popularity of political entities

Research question: *What was the temporal trend of support for two major political parties: BJP and Congress?*

We first consider two text templates: [vote for] and [will win]. Among 4,381,623 unique bigrams, [vote for] and [will win] rank 16^{th} and 269^{th}, respectively and are the top two bigrams that can be used to express political preference (candidate or party). The tokens that immediately follow/precede [vote for]/[will win] are visualized in Figure 3 with the two main takeaways:

- Narendra Modi and BJP had overwhelming support.



(a) vote for (b) will win

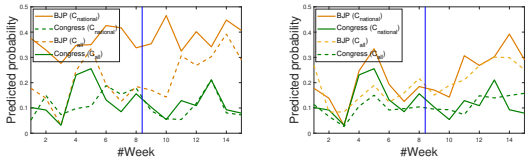
Figure 3: A word cloud visualization of [vote for] and [will win]. Among 4,381,623 unique bigrams, in terms of frequency, [vote for] and [will win] rank 16^{th} and 269^{th}, respectively.

- As compared to Narendra Modi and BJP, Rahul Gandhi and Congress had substantially less support.

We now present our results probing fine-tuned BERT. We constructed two cloze statements: [Vote for MASK] (denoted as $\mathcal{S}_3$) and [MASK will win] (denoted as $\mathcal{S}_4$). In our next series of experiments, we chose the granularity of weekly results. We divide the comments into weekly subsets based on the week they were posted yielding one corpus per week in the time-frame considered. Next, BERT was fine-tuned on each of these corpora yielding one fine-tuned BERT model per week. We queried each of these weekly fine-tuned BERT models with $\mathcal{S}_3$ and $\mathcal{S}_4$ and examined the results.

5.3.1 *Party-focused analysis*

For every week, among the ranked predictions, BJP and Congress consistently featured as the top-two political parties. We found this result consistent with the ground truth that indeed, these two parties are the two most-popular national parties. In Figure 4, we plot the predicted probabilities for BJP and Congress. As shown in Figure 4, both on $\mathcal{C}_{national}$ and $\mathcal{C}_{all}$, BJP was assigned a higher probability than Congress. In Figure 4(b), apparently, support for both parties showed a sharp decline in week 3. This week coincides with the period of heightened tensions between India and Pakistan [1] and a substantial chunk of the corpus discussed a potential war and possible outcomes. For the templates used for querying, the probabilities got split among India and Pakistan (in addition to the political entities) i.e. a substantial chunk of the users were discussing who would win a hypothetical war (India/Pakistan will win).



(a) [Vote for MASK] (b) [MASK will win]

Figure 4: Party focused analysis. BJP is plotted with saffron and Congress is plotted with green. Blue line indicates the time when voting starts. Solid lines indicate $\mathcal{C}_{national}$. Dotted lines indicate $\mathcal{C}_{all}$.

Robustness to phrase variation: One might argue that a simple frequentist analysis of plotting weekly occurrence `[vote for BJP]` or `[vote for Congress]` normalized by the total number of weekly occurrence of `[vote for]` can be equally effective in indicating BJP's dominance over Congress throughout the entire period. However, for less common phrases with similar meaning, lack of exact match can make this type of frequentist analysis difficult. For example, `[cast your vote to]` has sparse presence in the corpus (22 exact matches in the entire corpus) as compared to 26,301 mentions of `[vote for]`, indicating a simple template-based matching (i.e. executing an exact phrase-match against the comments) would not work (sophisticated embedding-based methods may address this issue). Querying a language model has an advantage for uncommon but similar meaning phrases as we could easily compute and compare probabilities with this template.

Comparison to Polls and Outcomes: Election laws in India ban the release of polling information close to an election. Exit polls are thus released after the election. The vast majority of the polls predicted a victory for the incumbent (with widely varying seat counts). This aligns with our discovery using the queries mentioned where the incumbent party, BJP, is assigned a higher probability than the opposition, Congress.

5.3.2 *Candidate-focused analysis*

We now move to our candidate-focused analysis. We used the same set of probes, $\mathcal{S}_3$ and $\mathcal{S}_4$ for our analysis and compared the two most-popular candidates: Narendra Modi (popularly referred to as Modi) and Rahul Gandhi (popularly, referred to as Rahul). As shown in Figure 5, Modi was overwhelmingly more popular than Rahul across the entire time-period we considered. This finding is consistent with previous finding [19] about 2014 elections and a Pew research survey [6] stating that 88% of the surveyed Indian citizens viewed him favorably. In contrast, the support for Rahul was very low. This finding is again consistent with the two following outcomes (i) Rahul lost in a seat that was held by Congress party and his family for years which was considered a party stronghold, and (ii) Rahul resigned as the party president following Congress's poor performance [5].

An adversarial example to highlight `BERT`'s robustness and Modi's overwhelming popularity: We have already shown that `BERT` is robust to phrase variations. We next show a stronger result. We remove any comment containing the phrase `[vote for Modi]` (or `[Modi will win]`) from the corpus and fine-tune `BERT` on the modified corpora. If `BERT` was only relying on counting statistics without forming a deeper understanding of the corpus, $\mathcal{S}_3$ and $\mathcal{S}_4$ should be completed with Rahul with higher probability than Modi. However, as shown in Table 9, Modi still received higher probability than Rahul indicating that `BERT` could still infer stronger support for Modi from the rest of the comments.

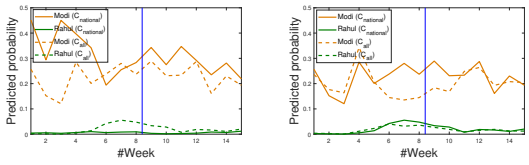
One user one comment: Fake accounts, bots, and a variety of manual or automated mechanisms exist to drive popularity or attention to entities. We re-ran all our analyses on a corpus where only one comment is randomly sampled and retained per user (similar to one person one vote) with qualitatively similar results. Note that, this does not eliminate the effects caused by multiple fake accounts detecting which is beyond the scope of this paper.

5.4 Deeper insights

Research question: *Is it possible to mine deeper insights like identifying the national priorities using `BERT`?*

Data set	Removed phrase	Probe	P(Modi)	P(Rahul)
$\mathcal{C}_{all}$	-	$\mathcal{S}_3$	0.2345	0.0065
$\mathcal{C}_{all}$	Vote for Modi	$\mathcal{S}_3$	0.1025	0.0042
$\mathcal{C}_{all}$	-	$\mathcal{S}_4$	0.2549	0.0219
$\mathcal{C}_{all}$	Modi will win	$\mathcal{S}_4$	0.1721	0.0122

Table 9: Performance on the adversarial corpora.



(a) `[Vote for MASK]` (b) `[MASK will win]`

Figure 5: Candidate focused analysis. Modi is plotted with saffron and Rahul is plotted with green. Blue line indicates the time when voting starts. Solid lines indicate $\mathcal{C}_{national}$. Dotted lines indicate $\mathcal{C}_{all}$.

So far, we have seen that querying `BERT` can be effective in (i) investigating sentiment around an entity (ii) comparing relative popularity between candidates and political parties and (iii) acting as a proxy for opinion/exit polls. In our next series of experiments, we explore if it is possible to obtain deeper insights. For this, we construct the following two cloze statements: `[The biggest problem of India is MASK]` (denoted as $\mathcal{S}_5$) and `[India's biggest problem is MASK]` (denoted as $\mathcal{S}_6$). For each month, we list top three predictions in Table 10.

In order to evaluate the effectiveness of `BERT` predictions, we would require a baseline ground truth to compare against. For this, we consider the most-recent survey conducted by Pew research [12] among 2,521 respondents in India from May 23 to July 23, 2018. Note that, there is a considerable time-lag between the conducted survey and our analysis during which a major terror attack happened in Pulwama which brought India and Pakistan almost to the brink of a full-fledged war. Hence, we observed some discrepancies between the survey's findings and `BERT` predictions. We attribute these to the highly significant and unexpected events that took place which explain the minor discrepancies.

The bag of problems identified by $\texttt{BERT}_{all}$ on $\mathcal{S}_5$ ($\{terrorism, corruption, Kashmir, unemployment, poverty\}$) and on $\mathcal{S}_6$ ($\{terrorism, Pakistan, Kashmir, corruption, unemployment, poverty, \}$) (see, Table 10) have substantial overlap indicating that the predictions are robust to simple phrase variations. Three issues identified with both cloze statements: terrorism, corruption and unemployment featured in the top four issues identified in the Pew research survey establishing that fine-tuning `BERT` on a massive web corpus can provide an interesting alternative to traditional surveys.

We next focus on the temporal nature of the predictions. While *terrorism* had a constant presence in the top three predictions from both cloze statements in all four months we considered, we notice that the predicted probabilities for *terrorism* was substantially higher in the month of February a time-period in which the Pulwama terror attack occurred. As the tensions between the two countries subsided, the other two pressing problems - corruption and unemployment started receiving more public attention. It is infeasible to conduct extensive field-surveys on a monthly basis. However, our results indicate that from a large data set of discussions on current events, it is possible to mine deeper insights and also analyze the temporal trends of public

Month	BERT$_{national}$ on $\mathcal{S}_5$	BERT$_{national}$ on $\mathcal{S}_6$	BERT$_{all}$ on $\mathcal{S}_5$	BERT$_{all}$ on $\mathcal{S}_6$
February	terrorism (0.24), Pakistan (0.17), corruption (0.14)	terrorism (0.49), Pakistan (0.09), corruption (0.04)	terrorism (0.28), corruption (0.16), Kashmir (0.06)	terrorism (0.37), Pakistan (0.13), kashmir (0.07)
March	unemployment (0.14) terrorism (0.09), corruption (0.09)	terrorism (0.20) Pakistan (0.09), Kashmir (0.05)	corruption (0.47), terrorism (0.12), poverty (0.10)	corruption (0.31), terrorism (0.30), poverty (0.05)
April	unemployment (0.29) poverty (0.14) corruption (0.07)	terrorism (0.12) unemployment (0.10) corruption (0.05)	corruption (0.36), unemployment (0.21), terrorism (0.07)	corruption (0.22), terrorism (0.15), Kashmir (0.07)
May	unemployment (0.21), corruption (0.19) terrorism (0.07)	terrorism (0.18), corruption (0.13) unemployment (0.05)	corruption (0.25), unemployment (0.21), poverty (0.08)	corruption (0.22), unemployment (0.14), terrorism (0.12)

Table 10: Predicted completions with probabilities in parentheses.

perception on national issues and priorities and LMs can provide a cost-effective, fast-turnaround alternative to traditional surveys.

Nested queries: We explore if we can go deeper and identify local issues through nested querying. In what follows, we show a preliminary study that holds promise. We first queried BERT$_{all}$ with the following cloze statement: `[MASK is a major city in Tamil Nadu]`. The result predicted with highest probability was Chennai. Next, we queried BERT$_{all}$, BERT$_{national}$, and BERT$_{TN}$, a BERT model fine-tuned only on subset of comments generated from Tamil YouTube channels with the cloze statements: `[Chennai's biggest problem is MASK]` and `[The biggest problem of Chennai is MASK]`. Our results show that while BERT$_{all}$ and BERT$_{national}$ both predicted *corruption*, *terrorism* and *unemployment*, the fine-tuned model specific to the state identified water crisis as one of the local issues. The Chennai water crisis [2] which started as a local issue snowballed into a national crisis that started receiving global attention in June, a time-frame beyond our analysis period. However, our results indicate that early detection of localized issue through focused analysis merits deeper exploration.

6 RETENTION OF PRIOR KNOWLEDGE

Owing to the black-box nature of large scale language models, it is unclear how fine-tuning impacts the existing knowledge in a model. In this section, we conduct what is to the best of our knowledge the first analysis of how knowledge from the original model is carried over to a fine-tuned model. We first re-iterate the following observations about our corpus:

1. Compared to typical training corpora used for the base models, our document lengths are considerably shorter.
2. The overlap of facts is fairly limited owing to the focus of the corpus.

Intuition suggests that most of the knowledge must stay intact or untouched by the fine-tuning step since the bulk of the corpus just deals with opinions about the Indian election and with entities relevant to this and other (smaller-scale) contemporaneous events in the Indian subcontinent.

We use the cloze sentences from [27] which cover a variety of domains such as entities and relations from ConceptNet, Google-RE, SQuAD. The sentences also cover a broad range of formats (querying for subjects, and objects; numeric literal values like year of birth etc.). BERT achieves reasonable performance on this corpus of questions and a strong argument is made for BERT's ability to serve as an open-domain Question-Answering (QA) model. Our experiment utilizes these very cloze sentences and by passing them as input to the BERT$_{base}$ (our base model) and our finetuned BERT$_{national}$, we are able to characterize the extent of knowledge lost during a fine-tuning step. We report P@1 scores and analyze the types of errors made by the fine-tuned model in Table 11.

Table 11 shows a slight decline in performance in all corpora. McNemar's test reveals these differences are statistically significant in most cases. We next analyze the types of errors introduced in the fine-tuned models and describe some patterns observed.

Corpus	Relation	#Facts	base	national	all
Google-RE	birth-place	2937	**40.17**	37.79	37.01
	death-place	1825	**24.57**	19.88	18.40
	birth-year	765	**3.34**	2.9	2.36
ConceptNet	Total	11458	**12.71**	10.55	10.85
SQuAD	Total	305	**13.11**	7.5	10.16

Table 11: P@1 performance of the cloze statements on the BERT$_{base}$ (base), BERT$_{national}$ (national), and BERT$_{all}$ (all). We observe a slight decline in performance across all corpora, and all types of relations.

Numerical Entities: Google-RE contains a set of cloze statements where the masked word is the year of birth of a subject (relation: birth-year in Table 11). We observed that in a lot of cases, the fine-tuned model predicted 1947 as the year of birth regardless of entity. 1947 is significant in Indian history (year of Independence from colonial rule). Our hypothesis is that year of birth has low support in the corpus (given that even the base model performs poorly) and thus it is trivial to move the distribution of numerical literals towards the distribution of years in the fine-tuning corpus.

Location Entities: In the birth-place and death-place relations, the decline in performance occurred due to a variety of mis-predictions. It is interesting to note that some of the mistakes were geographically close to the correct answer, for instance [Tehran to Iran], [Glasgow to London], [Hartford to Greenwich]. We did not however observe an over-representation of Indian locations in the result set. This is possibly due to the limited mention of cities in our corpus (verified by manual inspection).

7 CONCLUSIONS

In this paper, in the context of the 2019 Indian general election, we evaluate the viability of fine-tuned large-scale language models in navigating and mining insights from corpora. Our fine-tuned models when queried reveal a variety of insights like temporal trends of candidate popularity, evolving national priorities, concerns of a population and sentiment around religions. We demonstrate through carefully constructed experiments that language modeling is robust to sparsity of the phrases queried and can operate even in situations when template-matching would fail. We corroborate the mined insights with manual analyses involving word-cloud tools, lexicon sentiment analysis tools, political outcomes and available surveys. Further, using our corpus, we produce a quantitative evaluation of a fine-tuned model's retained knowledge, and provide insights about what is retained, acquired, and forgotten. We posit that improved language models of the future can provide a viable alternative to existing IR pipelines for analysis and mining.

REFERENCES

[1] Bbc. https://www.bbc.com/news/world-asia-47366718. Online; accessed 12-March-2019.

[2] Cnn. https://www.cnn.com/2019/07/12/india/india-chennai-water-crisis-train-intl/index.html. Online; accessed 16-Aug-2019.

[3] Election commission of india. https://eci.gov.in/about/about-eci/the-functions-electoral-system-of-india-r2/. Online; accessed 16-Aug-2019.

[4] New york times. https://www.nytimes.com/2019/04/11/world/asia/modi-india-elections.html. Online; accessed 28-July-2019.

[5] New york times. https://www.nytimes.com/2019/07/03/world/asia/rahul-gandhi-resigns.html. Online; accessed 12-March-2019.

[6] Pew research center. https://www.pewresearch.org/global/2017/11/15/india-modi-remains-very-popular-three-years-in/. Online; accessed 16-Aug-2019.

[7] Washington post. https://www.washingtonpost.com/world/asia_pacific/divided-families-and-tense-silences-us-style-polarization-arrives-in-india/2019/05/18/734bfdc6-5bb3-11e9-98d4-844088d135f2_story.html. Online; accessed 28-July-2019.

[8] RB Bhagat, 'Census enumeration, religious identity and communal polarization in india', *Asian Ethnicity*, **14**(4), 434–448, (2013).

[9] Aditya Bohra, Deepanshu Vijay, Vinay Singh, Syed Sarfaraz Akhtar, and Manish Shrivastava, 'A dataset of hindi-english code-mixed social media text for hate speech detection', in *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, pp. 36–41, (2018).

[10] Pete Burnap, Rachel Gibson, Luke Sloan, Rosalynd Southern, and Matthew Williams, '140 characters to victory?: Using twitter to predict the uk 2015 general election', *Electoral Studies*, **41**, 230–233, (2016).

[11] Fabio Celli, Evgeny Stepanov, Massimo Poesio, and Giuseppe Riccardi, 'Predicting brexit: Classifying agreement is better than sentiment and pollsters', in *Proceedings of the Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media (PEOPLES)*, pp. 110–118, (2016).

[12] Pew Research Center. A sampling of public opinion in india, 2019.

[13] Dorottya Demszky, Nikhil Garg, Rob Voigt, James Zou, Matthew Gentzkow, Jesse Shapiro, and Dan Jurafsky, 'Analyzing polarization in social media: Method and application to tweets on 21 mass shootings', in *Proceedings of the 17th Annual NAAC)*, (2019).

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, '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)*, pp. 4171–4186, (June 2019).

[15] Joseph DiGrazia, Karissa McKelvey, Johan Bollen, and Fabio Rojas, 'More tweets, more votes: Social media as a quantitative indicator of political behavior', *PloS one*, **8**(11), e79449, (2013).

[16] Asgharali Engineer, *Communal riots in post-independence India*, Universities Press, 1997.

[17] William L Hamilton, Kevin Clark, Jure Leskovec, and Dan Jurafsky, 'Inducing domain-specific sentiment lexicons from unlabeled corpora', in *Proceedings of EMNLP*, volume 2016, p. 595. NIH Public Access, (2016).

[18] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov, 'Bag of tricks for efficient text classification', *arXiv preprint arXiv:1607.01759*, (2016).

[19] Vadim Kagan, Andrew Stevens, and VS Subrahmanian, 'Using twitter sentiment to forecast the 2013 pakistani election and the 2014 indian election', *IEEE Intelligent Systems*, **30**(1), 2–5, (2015).

[20] Raghav Kapoor, Yaman Kumar, Kshitij Rajput, Rajiv Ratn Shah, Ponnurangam Kumaraguru, and Roger Zimmermann, 'Mind your language: Abuse and offense detection for code-switched languages', *arXiv preprint arXiv:1809.08652*, (2018).

[21] Nora Kassner and Hinrich Schütze, 'Negated lama: Birds cannot fly', *arXiv preprint arXiv:1911.03343*, (2019).

[22] Aparup Khatua, Apalak Khatua, Kuntal Ghosh, and Nabendu Chaki, 'Can# twitter_trends predict election results? evidence from 2014 indian general election', in *2015 48th Hawaii international conference on system sciences*, pp. 1676–1685. IEEE, (2015).

[23] Brendan O'Connor, Ramnath Balasubramanyan, Bryan R Routledge, and Noah A Smith, 'From tweets to polls: Linking text sentiment to public opinion time series', in *Fourth International AAAI Conference on Weblogs and Social Media*, (2010).

[24] Shriphani Palakodety, Ashiqur R. KhudaBukhsh, and Jaime G. Carbonell, 'Hope Speech Detection: A Computational Analysis of the Voice of Peace', *CoRR*, **abs/1909.12940**, (2019).

[25] Shriphani Palakodety, Ashiqur R. KhudaBukhsh, and Jaime G. Carbonell, 'Voice for the Voiceless: Active Sampling for Finding Comments Supporting the Rohingyas', in *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20)*, p. To appear, (2020).

[26] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer, 'Deep contextualized word representations', in *Proc. of NAACL*, (2018).

[27] Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller, 'Language models as knowledge bases?', 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)*, pp. 2463–2473, Hong Kong, China, (November 2019). Association for Computational Linguistics.

[28] Erik Tjong Kim Sang and Johan Bos, 'Predicting the 2011 dutch senate election results with twitter', in *Proceedings of the workshop on semantic analysis in social media*, pp. 53–60, (2012).

[29] NC Saxena, 'The nature and origin of communal riots in india', *Communal riots in post-independence India*, **60**, (1984).

[30] Parul Sharma and Teng-Sheng Moh, 'Prediction of indian election using sentiment analysis on hindi twitter', in *2016 IEEE International Conference on Big Data (Big Data)*, pp. 1966–1971. IEEE, (2016).

[31] Min Song, Meen Chul Kim, and Yoo Kyung Jeong, 'Analyzing the political landscape of 2012 korean presidential election in twitter', *IEEE Intelligent Systems*, **29**(2), 18–26, (2014).

[32] Chi Sun, Luyao Huang, and Xipeng Qiu, 'Utilizing bert for aspect-based sentiment analysis via constructing auxiliary sentence', *arXiv preprint arXiv:1903.09588*, (2019).

[33] Ian Talbot and Gurharpal Singh, *The partition of India*, Cambridge University Press Cambridge, 2009.

[34] Adam Tsakalidis, Nikolaos Aletras, Alexandra I Cristea, and Maria Liakata, 'Nowcasting the stance of social media users in a sudden vote: The case of the greek referendum', in *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pp. 367–376. ACM, (2018).

[35] Andranik Tumasjan, Timm O Sprenger, Philipp G Sandner, and Isabell M Welpe, 'Predicting elections with twitter: What 140 characters reveal about political sentiment', in *Fourth international AAAI conference on weblogs and social media*, (2010).

[36] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin, 'Attention is all you need', in *Advances in neural information processing systems*, pp. 5998–6008, (2017).

[37] Leonid Velikovich, Sasha Blair-Goldensohn, Kerry Hannan, and Ryan McDonald, 'The viability of web-derived polarity lexicons', in *NAACL*, pp. 777–785. Association for Computational Linguistics, (2010).

[38] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le, 'Xlnet: Generalized autoregressive pretraining for language understanding', *CoRR*, **abs/1906.08237**, (2019).