IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 2 February 2024, accepted 20 February 2024, date of publication 22 February 2024, date of current version 29 February 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3368929

RESEARCH ARTICLE

Textual Pre-Trained Models for Age Screening Across Community Question-Answering

ALEJANDRO FIGUEROA[1] **AND MOHAN TIMILSINA**[2]
[1]Departamento de Informática y Computación, Universidad Tecnológica Metropolitana, Santiago 7800002, Chile
[2]Data Science Institute, Insight Centre for Data Analytics, University of Galway, Galway, H91 TK33 Ireland

Corresponding author: Alejandro Figueroa (alejandro.figueroa@utem.cl)

This work was supported in part by the project Fondecyt "Multimodal Demographics and Psychographics for Improving Engagement in Question Answering Communities" funded by Chilean Government under Grant 1220367, and in part by the Patagón Supercomputer of Universidad Austral de Chile under Grant FONDEQUIP EQM180042.

ABSTRACT Almost every community Question-Answering (cQA) platform has the pressing need of enhancing user experience by presenting dedicated displays, connecting potential answerers with open questions and revitalizing the material in their archives. In doing so, it is crucial to understand the profile of their community members, especially as it relates to their demographics. In this realm, variables such as age and gender have shown to be particularly promising for managing content. For instance, they make it easier to connect questions posted by one generation that are more likely to be answered by individuals from the previous generation. This paper advances the current body of knowledge in this area by exploring the performance of nineteen frontier transformer-based models (e.g., BERT and ELECTRA) on age recognition across a large-scale collection of cQA members. In effect, the best encoder (LongFormer) finished with an accuracy of 78.61% (F1-Score of 0.7424) by taking full-questions and answers into account. Unlike gender recognition, our outcomes do not show a noticeable difference between cased and uncased models. But on the other hand, they confirm that the transition from one age group to the other is smooth, and thus boundary individuals pose a tough challenge to discriminant models built on top of frontier machine learning approaches.

INDEX TERMS Pre-trained models, user analysis, question answering, age demographics, transformers.

I. INTRODUCTION

Automatically identifying age across community members is pivotal for almost all types of online social networks. As a matter of fact, these services not only need this piece of information to enhance user experience, but also they have the utmost urgency to the enforcement of their terms of service and the corresponding local laws. In effect, age screening is vital for successfully detecting many malicious activities, e.g., identify theft and fake profile creation as well as protecting children from harmful situations.

In the event of cQAs, previous studies have demonstrated that determining age is regarded as a decisive factor in matching open questions with potential answerers, especially in platforms where there is a transfer of knowledge from one generation to the other [1]. More concretely, it has been discovered that questions prompted by one particular age group are more likely to be answered by, typically another, specific cohort. More distinctively when dealing with topics including Dining Out, Home & Garden, and Family & Relationships.

However, age screening is not as simple as it might sound since abusers, criminals, and young people get around it by misrepresenting their age. Nevertheless, recent advances in Artificial Intelligence, namely Natural Language Processing and network analysis, have revealed that it is possible to infer textual, visual and activity patterns that are representative of distinct age clusters [1], [2], [3]. Needless to say, image and graph-based pre-trained Deep Neural Networks have offered great help here, but on the other hand, text-based frontier

The associate editor coordinating the review of this manuscript and approving it for publication was Arianna Dulizia.

architectures have not been largely explored yet despite their significant recent breakthrough in this field [4], [5], [6], [7], [8], [9].

Hence, the novelty of this paper lies in comparing the performance of assorted state-of-the-art pre-trained models (PTMs) working on textual inputs. With this in mind, our contributions to this body of knowledge are summarized as follows:

1) We adapted nineteen different frontier transformers, capable of achieving high age classification rates on the basis of textual inputs only.
2) All our experiments are conducted on a massive study corpus, namely almost 550k community member profiles encompassing their respective full questions, answers and self-descriptions.
3) We provide empirical evidence that the transition from one age group to the next/previous is smooth, and thus it presents a difficult challenge, even to state-of-the-art machine learning classification techniques.

In a nutshell, the best transformer was LongFormer, which accomplished an accuracy of 78.61%. The roadmap of this work is as follows. First, previous studies are discussed in Section II, and later Section III outlines our research questions. Then, Section IV and V discuss our methodology and experimental results, respectively. Eventually, Section VI touches on the key findings and sketches some future research directions.

II. RELATED WORK

In fact, age demographics across cQA websites is a largely unexplored research area. By examining the impact of sentiment analysis on these services, [10] superficially explored age patterns, in particular as they relate to attitude and sentimentality. In this regard, the authors presented two key findings: a) people are likely to respond to people of the same age in a more positive manner; and b) the sentimentality decreases with increasing age.

In a similar spirit, [11] focused on age trends across programming gurus in StackOverflow. They learned that programmer reputation scores grow in tandem with age well into the 50's, while in their 30' they tend to focus on fewer areas wrt. younger or older members. Additionally, they did not notice a strong correlation between age and scores in any specific knowledge area.

Lately, [3] perceived age recognition as a classification task by trying several ways of clustering community peers in consonance with their birth year. Curiously enough, their outcomes unveiled that it is better to reduce the archetypal five generations proposed by Strauss and Howe [12] to three via grouping its oldest three cohorts into one.

Consistent with this previous study, [1] tested high-dimensional vector spaces constructed from textual and metadata properties on conventional statistical strategies and frontier deep neural networks. In summary, FastText and MaxEnt proven to be effective, and when it comes to features, sentiment analysis. But more importantly, they drew the conclusion that effective models for identifying age cohorts bear strikingly similarities to models previously designed for question intent [13], namely in relation to the pertinence of sentiment analysis. Later, [14] provided support for the validity of this hypotheses by comparing the classification rate of assorted single-task and multi-task frameworks. More recently, the study of [2] showed that age-based centroid vectors tend to form a trail ordered by age in graph-based embeddings constructed on top of the activity of community.

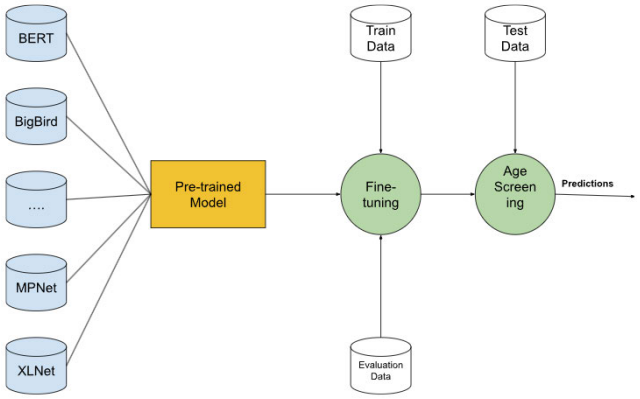


FIGURE 1. The flowchart of PTMs for age screening across cQA members. First, a pre-trained model is selected and downloaded from Hugging Face. This PTM is then adjusted to our task via the training and evaluation sets. The fine-tuned model is eventually utilized for casting predictions across the test collection.

III. RESEARCH QUESTIONS

With prior works as the foundation, we advance this area of research by answering the following two main research questions:

- **RQ1: How well do vanilla frontier neural network classifiers perform on age identification?**
- **RQ2: What can we learn from the errors that would help in the design of more efficient models in the future?**

IV. OUR METHODOLOGY

In this section, we outline the assorted pre-trained encoders employed in our empirical settings for recognizing age (see figure 1). BERT is one of the first and most widely used PTMs. And its name is the acronym for Bidirectional Encoder Representations from Transformers, (cf. [15], [16]). This architecture consists of a multi-layer bidirectional transformer trained on clean plain text for masked words and next sentence prediction [16], [17]. Its underlying idea is the old principle that words are, to at least a great extend, defined by other terms within the same context [18]. BERT is composed of twelve transformer blocks and twelve self-attention heads with a hidden state of 768. Our experimental configurations accounted for most state-of-the-art BERT-inspired architectures. Fundamentally, we extended

the encoders utilized by [19] (e.g., ALBERT [20], DEBERTA [21] and XLNet [22]) as described below:

- **BigBird** deals with the inherent quadratic dependency of BERT by implementing a sparse attention mechanism. This new attention approach is linear in the number of tokens [23]. This model capitalizes on the power of extra global tokens for preserving its expressiveness. As a result, this brings about competitive performance on downstream tasks like question answering and long document classification.
- **ELECTRA** (Efficiently Learning an Encoder that Classifies Token Replacements Accurately) adapts a discriminator (transformer) that determines whether every token is an original or a replacement, instead of only masking a fraction of tokens within the input [24]. A generator, another neural network, masks and substitutes tokens to generate corrupted samples. In practical terms, this model trains much faster than BERT, requiring significantly less computation, while at the same time, accomplishing a competitive accuracy on several downstream tasks. In this work, we accounted for its base and large versions.
- **MPNET**capitalizes on underlying dependencies among predicted tokens via permuted language modeling and takes additional position information as input. This means it can see a full sentence, and lessen the position discrepancy accordingly [25]. In other words, it leverages the advantages of BERT and XLNet, while at the same time avoiding their limitations.
- **SqueezeBERT** replaces several operations within self-attention layers by grouped convolutions. As a consequence, it is four times faster than BERT while accomplishing a competitive performance. This architecture also runs at lower latency on smartphones than many efficient encoders like MobileBERT [26].

On the whole, we took advantage of a total of 19 pre-trained models including DistillBERT [27] and RoBERTa [28] (see the complete list on table 3). For fine-tuning, we profited from the implementations by Hugging Face[1] using the Simple Transformers[2] library. By and large, we opted for default parameters settings to level the grounds and also to reduce the experimental workload. At all times, two epochs were set during model adjustment, this way the time for fine-tuning was limited to ten days (largest models). It is worth noting here that going beyond one epoch did not show any significant improvement, but we intentionally gave all encoders enough time to converge. The maximum sequence length was equal to 512, and whenever possible, we set sliding windows operating with a 0.95 stride. The reader can refer to table 4 for details about experimental configurations using and not using sliding windows. As for the batch size, this was set to make sure that the corresponding GPU memory usage reached its limit, namely eight due to the largest models (XLNet, XLMRoBERTa, DEBERTA, etc.), this way always allowing convergence. In our experiments, We used twenty NVIDIA Tesla GPU cards: 1 x P4 (8gb), 16 x A16 (16gb) and 3 x P40 (24gb). Lastly, it is worth mentioning that we applied half precision (fp16) format to all models, but FNET.

[1] huggingface.co
[2] github.com/ThilinaRajapakse/simpletransformers

V. EXPERIMENTS

In the first place, we capitalized on the collection used by [19] for undertaking their gender study across cQA platforms. This corpus was gathered by [3], and it comprises 548,375 cQA member profiles including demographic information such as gender and age. Each of these records is connected with its corresponding sets of questions, answers, nicknames and self-descriptions (see table 1). Following the suggestion of [3], we clustered community fellows in conformity to the three groups version of the segmentation proposed by Strauss and Howe [12] (see distribution on table 2). Accordingly, the Entropy of this demographic variable is 0.8477.

This set was randomly divided into training 329,025 (60%), evaluation 109,675 (20%), and testing 109,675 (20%) by means of a random stratified sampling strategy, maintaining similar class distributions across the three splits. Particularly, in the test set, we find 53,842 instances from Gen Z (49,09%), 45,997 from Gen Y (41,94%) and 9,836 from Olders (8,97%). It is worth highlighting here that held-out evaluations were conducted in all our experiments by keeping these splits unchanged. Also, it has to be clarified that test material was employed solely for providing an unbiased assessment of a final model fit on the training/evaluation datasets.

In our experiments, we accounted for four empirical variations with the aim of assessing the contribution of each individual kind of information within the user profile to the overall performance of classifier. These four configurations are signalled by the next abbreviations:

- T (question titles only)
- TB (questions titles plus question bodies)
- TBA (full questions coupled with answers)
- TBAD (full questions, answers and self-descriptions)

Tables 3 and 4 displays the outcomes accomplished by the different combinations of encoders and configurations. From these results, it worth remarking the following findings:

1) Regardless the metric, MobileBert finished with the best average performance. Note that every time it did not end with the best score, it provided a very competitive performance.
2) Contrary to pre-trained models for gender classification [19], self-descriptions have a positive impact on the average classification rate. Basically it enhanced the performance for many BERT and RoBERTa-based architectures, whereas diminishing the score of other encoders such as FNet. This uncertainty can be attributed to the sparseness of short bios, only 7% of the community fellows decribe themselves.

TABLE 1. Illustrative excerpts from three distinct profiles within our study corpus. Each sample belongs to a different age cohort. Italics denotes questions, whereas self-descriptions are underlined with a wavy line.

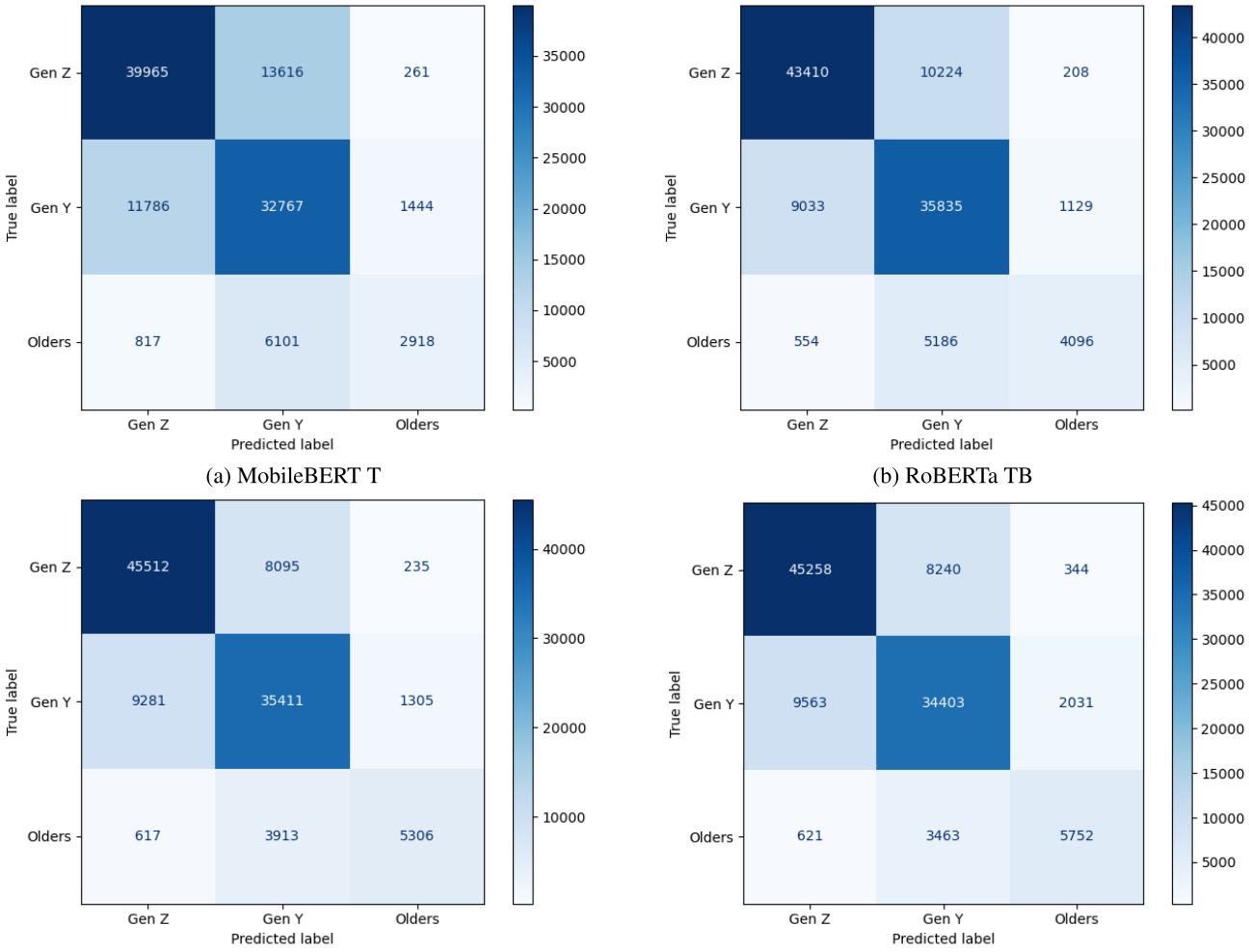
GEN Z
Do you think high school should be grades 8-12? The high schools where I live here in Canada are grades 8-12. Do you think it should be that way? Oh but I don't live in the US
Do you think it's okay for a 14 year old girl to wear these clothes? Do you think it's okay for a 14 year old to wear a black shirt that says, "I'm a whiskey, lipstick and tattoos kind of girl" that shows her bra and a pair of short shorts that might show her butt?
Who else thinks the songs "About The Bass" and "Bang Bang" are cringe worthy? I don't think those songs are cringe worthy because of what they're singing about (in Nicki Minaj's part in Bang Bang, it's rapping), they're just cringe worthy to me for some reason. I know tons of other songs that talk about similar things as those songs and a lot are really catchy and they don't cringe me..maybe I don't like the lyrics or the sound of the songs. Who else thinks they're cringe worthy????
Nope. I'm 14 but I would never consider doing that. I don't want my future possibly ruined. Maybe a crop top is good but naked pictures is a big no no.
Yes. Kids even younger than that now have beasts. You can look up a 10 year old child actress and see that she might be revealing her body and dressing up like a 20 year old.
i'm 14 and going to grade 9 :)
GEN Y
What would you do? I am almost 20, I am intimidated by driving b/c I'm always afraid I won't know where I'm going. Therefore I don't drive. Any suggestions how I can get over this?
Cheetahs can run up to 31.6 meters per second, while the world record speed of a man is 10.3 meters per second. Which translates to 60 miles per hour for the cheetah and roughly 20 miles per hour for the man. However a cheetah can only maintain its maximum speed for about two miles. I suppose if a man raced one after that the man would be faster, but otherwise the facts speak for themselves.
My mom makes like brocolli cheese rice with boiled chicken that she cuts and adds to the rice. We always eat it with applesauce. It takes 30 to 40 minutes to make but you don't have to use an oven, and its pretty easy you just follow the directions on the package. Also when you make the rice you have to add water, my mom uses the leftover water from the boiled chicken to add flavor.
I am a very introverted person. I tend to spend more time by myself than I do with other people, but that's the way I like it. I love my dogs, my family and I really enjoy reading, writing, word puzzles, knitting and surfing the web for fun and/or interesting games. I am currently a college student studying Child Development. If there's anything else you'd like to know, ask a question about it lol :P
Olders
I had to reboot from install discs and having reinstalled yahoo cannot get message box on compose mail HELPme!?
Of course the universe was simply NOT created in 10 days, try 14 billion years ... and take a look at this incredible site, it is amazing and wonderful http://www.postimage.org/image.php?v=gx1G8GWS And while on the subject, so far as those of us on this planet, we owe our daily life to the Sun (without it there is no life), we may be created by primitive if not primary life forms dumped here by comets and if we aren't a little more careful, we might just destroy it all. Pretty bloody dumb if we don't wake up an reverse our stupid ways, don't you think?
Do you understand why aboriginal people invented canoes, spears, woomeras even the brilliant boomerang? Little sad that our fantastic native people are so under-rated by so many who just accept that what "they" - meaning their ancestors in most cases - have "achieved" such as the wheel, the steam engine, airplanes, rockets, electricity, steel .... weapons, atomic bombs .... and in so doing helping to destroy the planet. The clue is in their incredible skills in hunting down the small amount of edible food. The answer can be found in it's beautiful expose by Jared Diamond in his "Guns, Germs and Steel". If you feel superior or that these wonderfully spirited people are lacking something and you are willing to open your mind to the wonderful truth, you can prove it by reading this book, otherwise, you need to close your mouth and not expose your ignorance or your racist attitudes.

TABLE 2. Definitions and distributions of the three age cohorts within our collection.

Cohort	Birth Years	Number of Samples
Gen Z	1995–2008	269,119 (49,07%)
Gen Y	1980–1994	229,631 (41,87%)
Olders	1910–1979	49,634 (9,05%)
Total		548,384

3) In juxtaposition, it is crystal clear that question bodies and answers gave a major boost to the best model. More concretely, its accuracy grew 7.02% when adding bodies to titles and 2.63% when enriching full questions with answers. A similar picture is seen in terms of F-Score, an increase from 0.6066 to 0.6953 when taking into account question content, and from 0.6953 to 0.7424 when also considering answers. All in all, informative cues inferred from answers are useful for building effective models for age recognition, in contrast to gender, where these inputs noticeably lower the classification rate [19].
4) Closely analogous to gender recognition [19], the worst performance can be, most of the time, attributed to: XLM-RoBERTa, BERT and ELECTRA. More specifically, XLM-RoBERTa reached the lowest scores when trained on full questions (70.25%) and on TBA (71.42%). We deem this as a result of their sensitivity to the distinct input signals.
5) The gap between the best and the worst systems is the narrowest when combining all signals (approximately 4.5% accuracy and 0.1103 F-Score points). Conversely, building models on top of full questions plus answers produces the widest gap (i.e., ca. 7.19% accuracy and 0.2135 F-Score points).
6) As for checking the statistical significance, we bootstrap sampled the best model for each configuration and each of its respective competitors twenty times and perform a two-tailed t-test afterwards ($\alpha < 0.05$). All but one pair passed this test: XLNet and RoBERTa (accuracy).

In summary, these outcomes point towards that age recognition is feasible on the basis of textual interactions within cQA platforms. Specifically, the best systems accomplished an accuracy of 78.76% (Longformer without self-descriptions) and a F-Score of 0.7676 (squeezebert-uncased with short bios). Results also point towards the fact that



(a) MobileBERT T (b) RoBERTa TB

(c) LongFormer TBA (d) MobileBERT TBAD

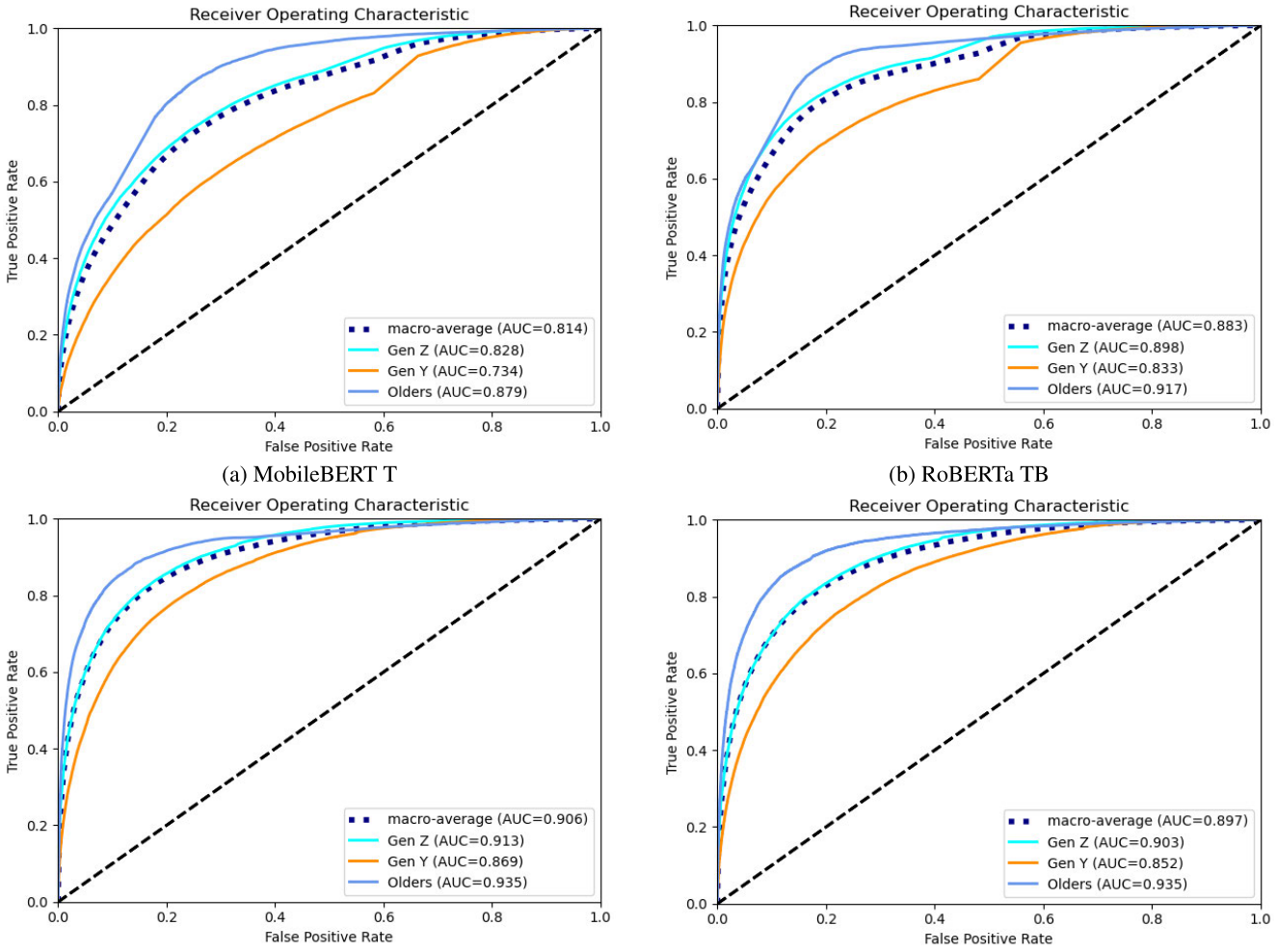
FIGURE 2. **The confusion matrix for the best model under each configuration.**

fine-tuning effective models for age recognition requires signals provided by answers, while prior studies proved that these are detrimental for gender [19]. On average, question bodies aided in enhancing the performance by 9.3%. By the same token, adding answers brought about an average increase of 3.02%, and subsequently including profile descriptions by 0.13%. In effect, our outcomes suggest that short bios might be omitted most of the times without significantly affecting the classification rate. Recall that these are very sparse, since only 7% of the profiles have self-descriptions.

Another empirical difference between age and gender detection lies in the benefits of casing, both case and uncased models achieve competitive performance in the event of age detection. To be more precise, MobileBert is uncased, whereas LongFormer and RoBERTa are cased. This result also entails that cost-efficient models, namely MobileBert, can rival more resource-demanding encoders (i.e., LongFormer and RoBERTa) similarly to gender [19].

Figure 2 displays the confusion matrices for the best encoder under each configuration. In the first place, results indicate that adding question bodies assisted in boosting the correct recognition of instances from all cohorts, while answers solely from GEN Z and Olders. Secondly, self-descriptions were profitable exclusively for improving the identification of the minority cluster, namely Olders. Roughly speaking, most informative cues about age can be found across full questions and by and large, answers and short descriptions are fruitful mainly for tackling the data sparseness caused by the minority cohort.

These four confusion matrices also corroborate the findings of previous works [1], [2], that is to say that errors arise from perceiving samples as members of the prior or the next cohort. In other words, few misclassifications occur due to conceiving GEN Z fellows as Olders or the other way around. Interestingly enough, our empirical outcomes support the discovery of [2], who found out that graph (community) activity patterns slowly and systematically change in tandem with age. Lastly, the precision markedly improved across the three



(a) MobileBERT T (b) RoBERTa TB

(c) LongFormer TBA (d) MobileBERT TBAD

FIGURE 3. ROC curves for the best model obtained for each configuration.

clusters, especially for the Olders. To be more exact, on GEN Z it went from 0.742 to 0.845 (compare figures 2a and 2c), while on GEN Y from 0.712 to 0.779 (see matrices 2a and 2b), and on Olders from 0.297 to 0.585 (cf. figs. 2a and 2d).

Additionally, figure 3 highlights the ROC (Receiver Operating Characteristic) curves for the best systems. These graphs show a progressive improvement from the titles only to the full questions plus answers setting. In fact, the macro-average AUC (Area Under the Curve) grew 0.092 points, and GEN Y had the sharpest increase (i.e., 0.135 points). All in all, self-descriptions were always detrimental in terms of AUC.

Figure 4 highlights term attributions[3] assigned by LongFormer TBA to three member profiles. In this image, positive attribution numbers (in green) indicate terms that positively contribute towards the predicted cohort, while negative numbers (in red) signal words that negatively contribute towards the recognized age group [29], [30]. A more intense shade of colour features a greater contribution. These scores reveal that terms such as ''academy'', ''music'' and ''dreams'' are strong indicators of the youngest generation, while ''parents'', ''permission'', ''mom'' and ''ask'' of GEN Y. In the event of Olders, we find words including ''benign'', ''vet'' and ''animals''. Interestingly enough, in this last group, terms like ''marriage'', ''woman'' and ''man'' contribute negatively. Since these words are normally found across the category ''Family and Relationships'', we conjecture that this negative contribution is due to its stronger relation to younger people, typically aiming or beginning to settle their families. All in all, this analysis shows that this best encoder reaped higher classification rates due to correctly inferring word usages within each target class.

Figure 5 unveils interesting regularities across errors. Like [1] noted, the accuracy systematically increases until around the middle of the Olders interval (i.e., year 1950), and it consistently decreases afterwards towards the boundary to GEN Y (year 1979). In a similar manner, the classification rate boosts from 1980 up to around the middle of the GEN Y

[3] github.com/cdpierse/transformers-interpret.git

TABLE 3. Accuracy obtained by each combination of transformer and pre-defined setting (test set). The † indicates that sliding windows could not be used.

Model	T	TB	TBA	TBAD	Max.	Avrg.	Min.	σ
albert-base-v1	0.6687	0.7299	0.7655	0.7661	0.7661	0.7326	0.6687	0.0458
bert-base-cased	0.6821	0.7157	0.7619	0.7338	0.7619	0.7234	0.6821	0.0334
bert-base-uncased	0.6732	0.7435	0.7681	0.7523	0.7681	0.7343	0.6732	0.0420
bigbird-base-trivia-itc	0.6785	0.7488	0.7719	0.7689	0.7719	0.7420	0.6785	0.0436
deberta-base	0.6839	0.7394	0.7208	0.7750	0.7705	0.7298	0.6839	0.0380
distilbert-base-cased	0.6802	0.7379	0.7686	0.7704	0.7704	0.7393	0.6802	0.0421
distilroberta-base	0.6841	0.7485	0.7747	0.7753	0.7753	0.7457	0.6841	0.0429
electra-base-discriminator	0.6623	0.7379	0.7612	0.7393	0.7612	0.7252	0.6623	0.0433
electra-small-discriminator	0.6626	0.7367	0.7459	0.7408	0.7459	0.7215	0.6626	0.0394
fnet-base	0.6828	0.7295†	0.7531†	0.7473†	0.7531	0.7282	0.6828	0.0319
longformer-base-4096	0.6760	0.7313	**0.7861**	0.7763	0.7861	0.7424	0.6760	0.0503
mpnet-base	0.6672†	0.7401†	0.7705†	0.7606†	0.7705	0.7346	0.6672	0.0467
mobilebert-uncased	**0.6896**	0.7515	0.7790†	**0.7788**†	0.7790	0.7497	0.6896	0.0421
roberta-base	0.6323	**0.7598**	0.7435	0.7468	0.7598	0.7206	0.6323	0.0593
squeezebert-mnli	0.6795	0.7354	0.7630†	0.7626†	0.7630	0.7351	0.6795	0.0393
squeezebert-mnli-headless	0.6800	0.7332	0.7627†	0.7624†	0.7627	0.7346	0.6800	0.0389
squeezebert-uncased	0.6821	0.7433	0.7692†	0.7676†	0.7692	0.7406	0.6821	0.0407
xlm-roberta-base	0.6785	0.7025	0.7142	0.7741	0.7741	0.7173	0.6785	0.0407
xlnet-base-cased	0.6835	0.7535	0.7622	0.7628	0.7628	0.7405	0.6835	0.0382
Max	0.6896	0.7598	0.7861	0.7788				
Avrg.	0.6751	0.7378	0.7601	0.7611				
Min	0.6323	0.7025	0.7142	0.7338				
σ	0.0128	0.0133	0.0182	0.0138				

TABLE 4. F1-Scores reaped by each encoder when dealing with the different pre-defined configurations (test set). The † signals that sliding windows could not be employed.

Model	T	TB	TBA	TBAD	Max.	Avrg.	Min.	σ
albert-base-v1	0.5781	0.6666	0.7207	0.7205	0.7207	0.6715	0.5781	0.0673
bert-base-cased	0.5976	0.6469	0.7181	0.6573	0.7181	0.6550	0.5976	0.0495
bert-base-uncased	0.5882	0.6741	0.7235	0.7050	0.7235	0.6727	0.5882	0.0599
bigbird-base-trivia-itc	0.5983	0.6782	0.7259	0.7244	0.7259	0.6817	0.5983	0.0598
deberta-base	0.6011	0.6781	0.5289	0.7340	0.7340	0.6355	0.5289	0.0896
distilbert-base-cased	0.5959	0.6688	0.7224	0.7215	0.7224	0.6772	0.5959	0.0597
distilroberta-base	0.5992	0.6832	0.7235	0.7251	0.7251	0.6828	0.5992	0.0590
electra-base-discriminator	0.5780	0.6727	0.6881	0.6975	0.6975	0.6591	0.5780	0.0550
electra-small-discriminator	0.5641	0.6624	0.6867	0.6744	0.6867	0.6469	0.5641	0.0561
fnet-base	0.5981	0.6588†	0.7006†	0.6972†	0.7006	0.6637	0.5981	0.0476
longformer-base-4096	0.5856	0.6621	**0.7424**	0.7350	0.7424	0.6813	0.5856	0.0734
mpnet-base	0.5382†	0.6728†	0.7271†	0.7155†	0.7271	0.6634	0.5382	0.0867
mobilebert-uncased	**0.6066**	0.6860	0.7400†	0.7386†	0.7400	0.6928	0.6066	0.0627
roberta-base	0.4422	**0.6953**	0.6969	0.7082	0.7082	0.6357	0.4422	0.1291
squeezebert-mnli	0.5850	0.6619	0.7156†	0.7131†	0.7156	0.6689	0.5850	0.0612
squeezebert-mnli-headless	0.5860	0.6600	0.7155†	0.7151†	0.7155	0.6692	0.5860	0.0613
squeezebert-uncased	0.5896	0.6726	0.7220†	**0.7676**†	0.7676	0.6880	0.5896	0.0762
xlm-roberta-base	0.5888	0.4986	0.6503	0.7166	0.7166	0.6136	0.4986	0.0927
xlnet-base-cased	0.6017	0.6886	0.7193	0.7217	0.7217	0.6828	0.6017	0.0561
Max	0.6066	0.6953	0.7424	0.7676				
Average	0.5801	0.6625	0.7036	0.7152				
Min	0.4422	0.4986	0.5289	0.6573				
Std. Dev.	0.0369	0.0414	0.0473	0.0237				

cohort (year 1987), and heavily falls thereafter until reaching the boundary to GEN Z. Subsequently, we see the steady rise and fall of the accuracy for instances from GEN Z (peaking at 2003). Given these results, we can conclude that misclassifications are more likely to be found around generational boundaries. This conclusion makes perfect sense since these individuals are ''transitional'', meaning it is reasonable to expect that they share considerable inter-generational traits, which are manifested across two consecutive cohorts. However, unlike [1], both inter-generational drops seem to be sharper in the case of LongFormer TBA. We interpret this as a consequence of inferring more informative cues of the right generation when dealing with ''transitional'' community peers. Another interesting discovery regards the rate of correct guesses, our outcomes indicate that this peaks around the middle of each age gap. We understand these to be the archetypes or the most representative users of the respective cohort. In effect, this finding is in consonance with the finding of [2], which revealed that age-based centroid vectors tend to form a trail ordered by age in

GEN Z

#s Do I Still have time to get into the Royal Academy Of Music In London ? Do I still have time to get into the Royal Academy Of Music ? I I have been dreaming of being able to get into the Royal Academy of Music but I ' ve heard it ' s really competitive and I ' m not too sure that I will be good enough to get in or should I just quit my dreams whilst I ' m young . I am self taught a the piano but I ' ve been playing since I was 9 and I can play grade 5 pieces . I am also only learning violin so I do n 't think I 'll be good enough to get in by then ? And I have grade 5 drums but I do n 't want really want to play drums though . Do you think with enough practice I could get into the Uni with any of these instruments ? Honest answers please . #/s

GEN Y

#s My ears were already pierced so i just gradually worked them in by putting tape in small amounts around regular ear rings and then slowly put them in . eventually they were big enough to fit the 8 gauge in . Over time , I moved up another 2 or 3 sizes , but a couple months ago i decided they were n 't for me and took them out . I did n 't ask my parents permission , but then again I have really laid back parents . My mom also thought they were gross , but she did n 't really care that I had them . so i really do n 't know what you should ask your mom , i just did it . but I ' m not trying to make that OK . If you feel you need her permission , and then she says no , you should respect that . you ' re not overweight at all . I ' m 5 ' 3 and 130 pounds . To me , that sounds perfectly healthy , and you are working out a lot . If anything , you need to make sure you do n 't loose too much weight . #/s

Olders

#s The Dirty Dancing soundtrack and The Forrest G ump soundtrack . They both have awesome classics on them , plus if you love the movies , they ' re extra bonus ! I love , that as a woman , I am not forced to have sex with government officials and I am not forced to be mut ilated in my genitals . I love that I have a chance to work hard to earn my money . I love that I can get an education and choose to have a career if I want . I love that no one forces me into a marriage with a man that I have never met before . about 3 or 4 I have a hard time hearing people who talk lower . I also have a hard time hearing when there is a lot of background noise . I mostly resort to trying to lip read . your elbow should be at 90 degrees warm bath For girls , lower abdomen , off to the side . for guys , upper back Orange Crush sprinkle c ayan pepper on them , one taste and the animals will flee y up ! i remember reading it in 5 th grade . defin et ly re cc om end it no , it has to do with your family history . ask your mom when she started and that will give you a good idea of when you will start Wicked ! ! ! go see a vet . it may be benign , so the vet may just do a small surgery and it could be taken care of and over . #/s

FIGURE 4. Accuracy vs. birth year (dotted line). The bars denote the fraction of community members born the respective year.

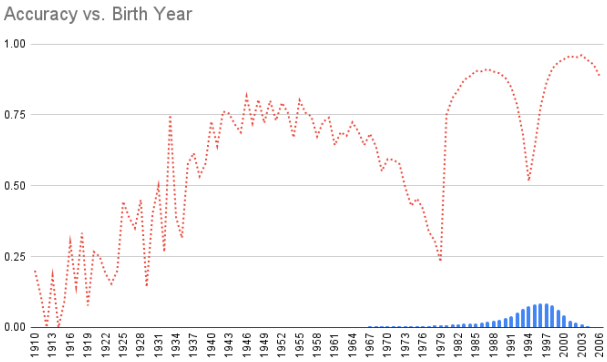


FIGURE 5. Accuracy vs. birth year (dotted line). The bars denote the fraction of community members born the respective year.

the Node2Vec embedding space built on top of activity graphs.

Additionally, figure 5 points towards the fact that the transition between GEN Y to GEN Z is smoother (i.e., substantially higher accuracy) than between Olders and GEN Y. This leads to the conclusion that the difference between GEN Y and GEN Z is wider than between Olders and GEN Y. Lastly, figure 5 discloses that the inter-generational decrease between GEN Y to GEN Z takes place despite of owning the largest fractions of samples, signalling that this drop is not due to data sparseness.

A. SUMMARY AND LIMITATIONS

Aside from previously mentioned considerations, we can see that short bios suffer severely from data-sparseness, since ca. 7% of community fellows entered their self-description on the Yahoo! Answers platform. Nevertheless, the real impact of these pieces of information must be still assessed. One can reasonably envision here that there is a good chance of extracting demographic nuggets from these contexts. In particular, on communities where their members might be more inclined to provide their short biographies, e.g., Quora, Reddit and Stack Exchange. Here, frontier transfer learning approaches can be employed to mitigate the data sparseness across Yahoo! Answers self-descriptions.

One limitation of our approach is that the transition from one age group to the other is smooth, this means conventional text-based frontier models have, at their best, insufficient input to deduce discriminative patterns for "transitional" individuals. Here it would make sense to extend this work to consider multi-modal sources of input signals such as their corresponding profile images and activity patterns.

Another limitation of this study regards the corpora on which these state-of-the-art encoders where pre-trained. Most of the times, this training corpora comprises web texts, Wikipedia and books. If there is significant computational power accessible, one could think about pre-training frontier transformers on massive amounts of user-generated cQA texts. We envisage that smaller architectures would be more cost-efficient, since these archives are not as big as vanilla collections of web texts, for instance. Still yet, this sort of pre-training presents interesting challenges, for example when it comes to misspellings, jargon, aliases and the cleaning of the corpora. Given the fact that the grammar across questions titles is sharply different to what we can find across question bodies and answers, we need to think about special adjustments or separate models.

Additionally, a promising way of dealing with community members who prompted few questions and/or yielded a low number of responses in English is by capitalizing on multi-lingual transformers. It seems perfectly logical that this might enhance the age recognition when users can express themselves in several languages. A good example of this are short bios written in a language different from English.

VI. CONCLUSION

In the first place, empirical results indicate that it is feasible to build frontier transformer-based classification models for age detection across cQAs. In this regard, LongFormer outclassed all other architectures when fine-tuned on all input signals but self-descriptions.

Unlike gender recognition, our outcomes do not show a noticeable difference between cased and uncased models. To be more precise, we see cased encoders, including RoBERTa and LongFormer, competing head-to-head with uncased MobileBERT.

We envision that our findings might be useful for transferring models to platforms where it is hard to obtain large-scale annotated data. Take for instance, services such as Stack Exchange, where people seldom tell cues about their age. We also envisage the implementation of multi-modal architectures as a means of obtaining a sharper detection within boundary samples.

Lastly, age recognition based on textual inputs is positioned to be an instrumental tool to design interventions

to promote equal engagement and participation in online communities.

REFERENCES

[1] A. Figueroa and M. Timilsina, "What identifies different age cohorts in yahoo! Answers?" *Knowl.-Based Syst.*, vol. 228, Sep. 2021, Art. no. 107278.
[2] M. Timilsina and A. Figueroa, "Neural age screening on question answering communities," *Eng. Appl. Artif. Intell.*, vol. 123, Aug. 2023, Art. no. 106219.
[3] A. Figueroa, B. Peralta, and O. Nicolis, "Coming to grips with age prediction on imbalanced multimodal community question answering data," *Information*, vol. 12, no. 2, p. 48, Jan. 2021.
[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. NIPS*, 2017, pp. 6000–6010.
[5] T. Lin, Y. Wang, X. Liu, and X. Qiu, "A survey of transformers," 2021, *arXiv:2106.04554*.
[6] C. Li, R. Cong, S. Kwong, J. Hou, H. Fu, G. Zhu, D. Zhang, and Q. Huang, "ASIF-Net: Attention steered interweave fusion network for RGB-D salient object detection," *IEEE Trans. Cybern.*, vol. 51, no. 1, pp. 88–100, Jan. 2021.
[7] W. Zhang, X. Sun, L. Zhou, X. Xie, W. Zhao, Z. Liang, and P. Zhuang, "Dual-branch collaborative learning network for crop disease identification," *Frontiers Plant Sci.*, vol. 14, Feb. 2023, Art. no. 1117478.
[8] W. Zhang, W. Zhao, J. Li, P. Zhuang, H. Sun, Y. Xu, and C. Li, "CVANet: Cascaded visual attention network for single image super-resolution," *Neural Netw.*, vol. 170, pp. 622–634, Feb. 2024.
[9] W. Zhang, Z. Li, G. Li, P. Zhuang, G. Hou, Q. Zhang, and C. Li, "GACNet: Generate adversarial-driven cross-aware network for hyperspectral wheat variety identification," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 3347745.
[10] O. Kucuktunc, B. B. Cambazoglu, I. Weber, and H. Ferhatosmanoglu, "A large-scale sentiment analysis for yahoo! Answers," in *Proc. 5th ACM Int. Conf. Web Search Data Mining*, New York, NY, USA, Feb. 2012, pp. 633–642.
[11] P. Morrison and E. Murphy-Hill, "Is programming knowledge related to age? An exploration of stack overflow," in *Proc. 10th Work. Conf. Mining Softw. Repositories (MSR)*, May 2013, pp. 69–72.
[12] B. Strauss, W. Strauss, and N. Howe, *Generations: The History of America's Future, 1584 to 2069*, 1991.
[13] D. Palomera and A. Figueroa, "Leveraging linguistic traits and semi-supervised learning to single out informational content across how-to community question-answering archives," *Inf. Sci.*, vol. 381, pp. 20–32, Mar. 2017.
[14] O. Díaz and A. Figueroa, "Improving question intent identification by exploiting its synergy with user age," *IEEE Access*, vol. 11, pp. 112044–112059, 2023.
[15] A. Radford and K. Narasimhan, *Improving Language Understanding By Generative Pre-training*, 2018.
[16] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.* Minneapolis, MI, USA: Association for Computational Linguistics, vol. 1, Jun. 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423
[17] C. Sun, X. Qiu, Y. Xu, and X. Huang, "How to fine-tune BERT for text classification?" 2019, *arXiv:1905.05583*.
[18] W. L. Taylor, "'Cloze procedure': A new tool for measuring readability," *Journalism Quart.*, vol. 30, no. 4, pp. 415–433, Sep. 1953.
[19] P. Schwarzenberg and A. Figueroa, "Textual pre-trained models for gender identification across community question-answering members," *IEEE Access*, vol. 11, pp. 3983–3995, 2023.
[20] P.-H. Chi, P.-H. Chung, T.-H. Wu, C.-C. Hsieh, Y.-H. Chen, S.-W. Li, and H.-Y. Lee, "Audio albert: A lite BERT for self-supervised learning of audio representation," in *Proc. IEEE Spoken Lang. Technol. Workshop (SLT)*, Addis Ababa, Ethiopia, Jan. 2021, pp. 344–350.
[21] P. He, X. Liu, J. Gao, and W. Chen, "DeBERTa: Decoding-enhanced BERT with disentangled attention," 2020, *arXiv:2006.03654*.
[22] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, "XLNet: Generalized autoregressive pretraining for language understanding," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–11.
[23] M. Zaheer, G. Guruganesh, A. Dubey, J. Ainslie, C. Alberti, S. Ontanon, P. Pham, A. Ravula, Q. Wang, L. Yang, and A. Ahmed, "Big bird: Transformers for longer sequences," 2020, *arXiv:2007.14062*.
[24] K. Clark, M.-T. Luong, Q. V. Le, and C. D. Manning, "ELECTRA: Pre-training text encoders as discriminators rather than generators," 2020, *arXiv:2003.10555*.
[25] K. Song, X. Tan, T. Qin, J. Lu, and T.-Y. Liu, "MPNet: Masked and permuted pre-training for language understanding," 2020, *arXiv:2004.09297*.
[26] F. Iandola, A. Shaw, R. Krishna, and K. Keutzer, "SqueezeBERT: What can computer vision teach NLP about efficient neural networks?" in *Proc. SustaiNLP, Workshop Simple Efficient Natural Lang. Process.*, 2020, pp. 124–135.
[27] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter," 2019, *arXiv:1910.01108*.
[28] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," 2019, *arXiv:1907.11692*.
[29] J. Vig, "A multiscale visualization of attention in the transformer model," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics, Syst. Demonstrations*, Florence, Italy, 2019, pp. 37–42.
[30] S. Abnar and W. Zuidema, "Quantifying attention flow in transformers," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 4190–4197.

ALEJANDRO FIGUEROA received the Ph.D. degree in computational linguistics from Universität des Saarlandes, Saarbrücken, Germany. He is currently an Associate Professor with the Department of Informatics and Computer Science, Universidad Tecnológica Metropolitana, Santiago, Chile. His research interests include natural language processing, machine learning, context grounding and multi-modality in question-answering systems, and information retrieval.

MOHAN TIMILSINA received the Ph.D. degree in computer science from the Data Science Institute, University of Galway, Ireland, in 2020. He is currently a Senior Postdoctoral Researcher with the University of Galway. His research interests include applied machine learning, bioinformatics, graph mining, and information retrieval from tabular and networked data.