IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 15 December 2022, accepted 2 January 2023, date of publication 10 January 2023, date of current version 13 January 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3235735

RESEARCH ARTICLE

Textual Pre-Trained Models for Gender Identification Across Community Question-Answering Members

PABLO SCHWARZENBERG[1], (Member, IEEE), AND ALEJANDRO FIGUEROA[2]

[1]Facultad de Ingeniería, Universidad Andrés Bello, Santiago 8370146, Chile
[2]Departamento de Ciencias de la Ingeniería, Facultad de Ingeniería, Universidad Andrés Bello, Santiago 8370146, Chile

Corresponding author: Pablo Schwarzenberg (pablo.schwarzenberg@unab.cl)

This work was supported in part by the Project Fondecyt "Multimodal Demographics and Psychographics for Improving Engagement in Question Answering Communities" by the Chilean Government under Grant 1220367; in part by the Patagón Supercomputer of the Universidad Austral de Chile under Grant FONDEQUIP EQM180042; and in part by the Center for Bioinformatics and Integrative Biology (CBIB), hosted at the Faculty of Biological Sciences, University Andrés Bello, Santiago, Chile.

ABSTRACT Promoting engagement and participation is vital for online social networks such as community Question-Answering (cQA) sites. One way of increasing the contribution of their members is by connecting their content with the right target audience. To achieve this goal, demographic analysis is pivotal in deciphering the interest of each community fellow. Indeed, demographic factors such as gender are fundamental in reducing the gender disparity across distinct topics. This work assesses the classification rate of assorted state-of-the-art transformer-based models (e.g., BERT and FNET) on the task of gender identification across cQA fellows. For this purpose, it benefited from a massive text-oriented corpus encompassing 548,375 member profiles including their respective full-questions, answers and self-descriptions. This assisted in conducting large-scale experiments considering distinct combinations of encoders and sources. Contrary to our initial intuition, in average terms, self-descriptions were detrimental due to their sparseness. In effect, the best transformer models achieved an AUC of 0.92 by taking full-questions and answers into account (i.e., DeBERTa and MobileBERT). Our qualitative results reveal that fine-tuning on user-generated content is affected by pre-training on clean corpora, and that this adverse effect can be mitigated by correcting the case of words.

INDEX TERMS Gender identification, community question-answering sites, engagement and participation in online communities, transformers.

I. INTRODUCTION

The term demography is universally understood as the study of human populations and their changes. It seeks to describe people in relation to characteristics, such as gender, age and religion. Therefore, demographic analysis is vital for identifying audiences and adapting content to their interests, levels of understanding, attitudes, and beliefs. In the case of cQA platforms, an audience-centered approach is crucial for maintaining an engaged community. It assists not only in encouraging increased participation by delivering attractive and targeted content according to personalized interests and motivations, but also in establishing effective connections between recently asked questions and community peers that can produce appropriate and timely responses. Intuitively, one form of achieving this is by designing landing/home pages tailored to each specific demographic segment and personality type.

Along with this, as might be expected, having easy access to demographic variables is useful to detect identity theft, fraud, to enforce terms of service and local laws, filtering and banning fake profiles. Simply put, these factors are particularly useful for properly dealing with assorted malicious activities. Incidentally, cQA sites also suffer from gender

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

differences since they tend to reflect our daily lives. One way these differences manifest is in the disparities in the number of female and male authors across their distinct topics [1]. Gender analysis plays a pivotal role in ensuring to their members the opportunity of a fair gender representation in categories with biased participation.

Like most of the websites that require membership, online social networks ask their newcomers to fill out a form with their personal information, when registering. In these forms, fields such as age and gender are optional. To a great degree, people choose ''rather not say'' due to discretion and/or they just want to get through the registration process as fast as they can. To help with user profiling, the last two decades of advances in machine learning and Natural Language Processing (NLP) have made it possible to infer informative patterns from textual content.

In the last couple of years, transformers have aroused intense interest due to their effectiveness in language understanding, vision and reinforcement learning [2], [3]. Consequently, extensive research has been undertaken to improve this class of models over the past few years in terms of its adaptation, efficiency and generalization. As a result, a rich variety of architectures currently exists, some of which have been devised to work well under certain conditions and to target specific tasks.

In short, this work enhances the existing body of knowledge on cQA platforms by assessing assorted state-of-the-art encoders for text-based gender recognition. More precisely, our study makes the following contributions:

1) Fine-tuning state-of-the-art pre-trained models, capable of gender identification from writing on cQA sites.
2) By benefiting from a massive automatically annotated dataset, we conduct a comprehensive empirical assessment of a wide variety of pre-trained transformers.
3) Experimental evidence showing that dataset similarity between the pre-trained architecture and downstream task influences the outcomes. Via NLP processing, transfer learning can be improved by updating the target dataset to increase its similarity with the dataset utilized for pre-training the encoder.

Our results suggest that each gender has its own distinctive patterns of interaction within cQA platforms, and that most of these differences are expressed in a way that is recognizable using natural language understanding techniques.

The remainder of this paper is organized as follows. Section II discuss related works. Sections III and IV present the research questions and methods, respectively. Section V discusses the experiments, results and findings. Finally, Sections VI and VII draw conclusions, limitations and outline future work.

II. RELATED WORK

First and foremost, the primary goal of this study is to compare the performance of various state-of-the-art transformers for automatically recognizing genders across cQA users. Due to the multifariousness of the human behavior, community fellows engage with these platforms in different ways and exhibit varying levels of activity. For instance, some participants use the site to ask questions and others interact with the site mainly to answer questions. Therefore, this work fine-tunes and assesses frontier encoders on several combinations of textual inputs, namely questions, answers and self-descriptions.

A. PRE-TRAINED DEEP NEURAL NETWORKS

The latest developments in neural networks allow the training of very deep architectures that can adequately cope with a vast variety of NLP tasks, such as text classification and machine translation. Beyond a shadow of a doubt, transformer models represent a significant breakthrough in this field [4], [5]. Their underlying idea has proven simple but very powerful. It consists of pre-training language model objectives on large networks with massive amounts of unlabeled data, and adjusting these networks to downstream tasks afterwards [6], [7]. OpenAI GPT and BERT are two pioneers of this approach [8], [9]. Since their inception, new variants have been devised to improve this first generation of encoders from different perspectives, including their adaptability, efficiency and generalization [10], [11], [12], [13].

Although these pre-trained models (PTMs) have achieved promising results in numerous difficult tasks, and thus turned into the ipso facto architecture for NLP [10], they still face many challenges: designing effective architectures, utilizing rich contexts, improving computational efficiency, and conducting interpretation and theoretical analysis [14]. It is an accepted fact that PTMs represent knowledge as real-valued vectors in contrast to symbolisms used by human beings.

It has been discovered that architectures such as BERT, capture linear word order and phrase-level information in their lower layers [15]. In particular, deeper tiers are needed to model long-distance dependencies (e.g., subject-verb agreements) [16]. Attention weights have shown to be weak indicators of subject-verb agreements and reflexive anaphora [15]. While there is a wide consensus in studies with different tasks, datasets, and methodologies that syntactic information is most prominent in the middle layers [17], there are some disagreements regarding semantic features. Some studies suggest that semantic features are encoded at the top, whereas others suggest that throughout the entire model [18]. In juxtaposition, surface features are codified at the bottom. Essentially, these models have been observed to imitate traditional tree structures [16] to represent the steps of the traditional NLP pipeline [18]. However, it is yet to be seen how well these findings transfer to domains with higher variability in syntactic structures (e.g., noisy user-generated content) and/or with more flexible word orders, as in morphologically richer languages [16].

Despite enabling important breakthroughs in various conventional NLP benchmarks, an increasing number of studies are revealing that their language skills are not as impressive as initially thought [17]. For example, it has been demonstrated

that they depend on shallow heuristics when classifying texts [19]. Although it is true that large PTMs are capable of holding a vast amount of knowledge, they typically fail if any reasoning is required on top of their stored facts [20], [21]. Moreover, some of this knowledge is lost after fine-tuning because of network capacity or under-representation of probing facts. Therefore, forgetting is not necessarily or significantly lessened by capitalizing on additional information harvested from larger corpora [22].

B. GENDER IDENTIFICATION ON cQA PLATFORMS

There are only a few studies addressing the detection of genders across cQA users, as evidenced by numerous recent surveys in this area [23], [24], [25], [26], [27]. Most of these studies relate to image processing, more specifically, to learn gender-informative visual patterns from profile avatars. For instance, heuristic methods have been utilized for automatically guessing genders on Stack Overflow [28]. Here, non-facial avatars pose a tough challenge even for ocular inspections [28]. Therefore, image-based pre-trained neural network models have also been evaluated using multifarious profile pictures [29].

On the opposite side of the spectrum, the research of [1] address the problem of automatically discriminating the gender of who asked a question using the question texts and metadata, demographics, and web searches. By building a wide diversity of high-dimensional vector spaces and exploiting the genders entered when the user signed up, they trained three supervised approaches on top of a large-scale corpora. They discovered that age, industry and second-level question categories were salient features of gender of an asker. Interestingly, the best text-only models sought to infer the same characteristics from semantic and dependency analyses.

On Yahoo! Answers, the investigation of [30] found some relationships between gender demographics and sentiment analysis, namely its synergy with attitude (i.e., inclination towards positive or negative sentiments) and sentimentality (i.e., number of sentiments). Women and men exhibited different attitudes across prompted questions and given answers: males were more neutral, whereas women were more positive in their questions and responses and were more sentimental when answering questions. Some gender differences across question types were found by [31] using data from the graphic design community on Stack Exchange and Quora. Women are more likely to respond questions seeking for opinions, while men produce more answers to factual questions on Stack Exchange. At both sites, responses from men had a more negative tone than women's answers, although this difference was not statistically significant.

Gender information cooperates in reducing the male-female inequality as it relates to their participation across distinct cQA categories. In this regard, it has been reported that females, who encounter other members of the same gender, are more likely to engage sooner than those who do not in Stack Overflow [32]. Another significant discovery discloses a stronger tendency among women to post more questions, whereas males to yield more answers, resulting in fewer thumb-ups for them, giving raise to lower average reputation scores for females [33], [34], [35]. Working under these findings, they designed a reputation strategy to lessen the gender gap that rewards points for publishing questions and answers to the same level. Along the same lines, the research conducted by [36] revealed that feminine users receive lower scores when responding, despite exhibiting higher efforts in their contributions, revealing some gender bias in the scoring of answers on sites like Stack Overflow. This bias, combined with the fact that gamification strategies such as scores and badges are more appealing to men than to women [33], supports the need to devise alternative strategies to promote women's participation in cQA sites, especially when anonymity is allowed, and gender information is not available.

Overall, recent studies point towards automatic gender identification as strategically vital to keep community members engaged with cQA websites.

III. RESEARCH QUESTIONS

By leveraging the power of transfer learning, we quantify and juxtapose the classification rate of assorted frontier pre-trained models, when fined-tuned for text-based gender detection. To this end, we analyzed the performance of these state-of-the-art encoders, by considering distinct combinations of the different textual contents found across member profiles (i.e., question titles and bodies, answers and self-descriptions).

Essentially, our predecessors have dealt with this subject by conducting analyses at the level of isolated questions only [1], or targeting profile avatars [29]. In this work, we extend this notion to all texts within his/her profile, that is, to consider all questions posted by the same community peer together with all his/her answers and self-descriptions.

Specifically, our primary goal is answering the following three research questions:

- **RQ1:** Is it possible to automatically detect gender across cQA members based on their textual interactions within the cQA site?
- **RQ2:** Are there any key differences in the performance among distinct encoders using similar input signals?
- **RQ3:** Are there any differences in the performance of the same model using different information?
- **RQ4:** What are the factors that influence the results obtained by the models?

IV. METHODOLOGY

In essence, our primary aim was to analyze and compare the performance of assorted PTMs on the task of automatic gender recognition on cQA websites.

One of the pioneers, and at the same time, one of the most widely used architectures is **BERT** (Bidirectional Encoder Representations from Transformers) [8], [9]. It is based on a multi-layer bidirectional transformer, trained on clean plain

text (i.e., the English Wikipedia and the BookCorpus) for masked words and next sentence prediction [6], [9]. BERT is able to understand the meaning of any word within a sentence in relation to "the company it keeps" [37], that is to say, all the remaining terms embodied within the same context. Its architecture consists of twelve transformer blocks and twelve self-attention heads with a hidden state of 768. To classify textual content, it represents an entire sequence by the final hidden state h of its first token [CLS]. Then, a softmax classifier is appended to its top as a means of predicting the odds of a category.

Thus, BERT has been a source of inspiration for many other architectures. From this perspective, we considered the most representative models in our empirical settings. These are briefly described below:

- **ALBERT (A Lite BERT)** modifies his predecessor in two substantial ways: a factorized embedding parametrization and it introduces a strategy for sharing cross-layer parameters [38]. The former facilitates the growth of the hidden size without markedly increasing the parameter number of the vocabulary embedding. The latter thwarts the number of parameters to escalate in tandem with the depth of the network. Both proposals reduce the memory consumption and the training time of BERT.
- **DeBERTa (Decoding-enhanced BERT with disentangled attention)** represents words via a vector that encodes their content and another vector its position. In addition, attention weights among terms are computed using disentangled matrices on their contents and relative positions. To predict masked tokens during pre-training, a mask decoder is utilized instead of an output softmax layer to incorporate absolute positions in the decoding layer. Furthermore, a new virtual adversarial training method were used for fine-tuning to improve generalization on downstream tasks [39].
- **DistilBERT** leverages knowledge distillation during pre-training to reduce the size of BERT, while maintaining almost all its language understanding capabilities. By using a triple loss, this reduction makes this model 60% faster, and through distillation via the supervision of a larger transformer, it is competitive on many downstream tasks [40].
- **DistilRoBERTa** is a distilled version of RoBERTa-base, obtained by training the model as DistilBERT. It has six layers, 768 dimensions, and twelve heads, decreasing the number of parameters from 125 to 82 million. On average, it is twice as fast as its predecessor.
- **ELECTRA (Efficiently Learning an Encoder that Classifies Token Replacements Accurately)** pre-trains 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 [41]. 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 compute, while at the same time, accomplishing a competitive accuracy on several downstream tasks.
- **FNET** replaces self-attention sub-layers with a simple unparameterized Fourier Transform on input tokens. It rivals efficient encoders while being much faster and lighter in memory demands. Because of its speed, the Fourier Transform demonstrated to be an efficient mixing mechanism [42].
- **Longformer** tackles the quadratic explosion caused by self-attention, when increasing the sequence lengths. As a substitute, his attention mechanism scales linearly via a drop-in replacement that amalgamates a locally windowed attention with a task motivated global attention, making it easier to process much longer documents [43].
- **MobileBERT** is a thin version of BERT that is equipped with bottleneck structures and a carefully designed balance between self-attentions and feed-forward networks. It trains a specially designed teacher model, an inverted-bottleneck incorporated BERT model that enables effective layer-wise progressive knowledge transfer [44].
- **RoBERTa** is a robust strategy for training BERT models [45]. In short, it uses longer training times and sequences, bigger batches, and one order of magnitude more data than BERT for training. This battery of design choices additionally removes the goal of guessing the next sentence and dynamically changes the masking pattern employed on the training data.
- **XLNet** is a generalized autoregressive pre-training strategy that learns bidirectional contexts. Instead of exploiting a fixed forward or backward factorization order, it maximizes the expected likelihood over all permutations of the factorization order [46].
- **XLM RoBERTa** is a transformer-based multilingual masked language model that is pre-trained on texts harvested from one hundred languages. This encoder achieves state-of-the-art performance in cross-lingual classification, sequence labeling and question answering, and strong improvements have been observed when coping with low-resource languages. Interestingly, these outcomes have been achieved while remaining competitive with frontier monolingual models [47].

We benefited from the pre-trained models supplied by Hugging Face[1] as detailed in table 1. We utilized the Simple Transformers[2] library for fine-tuning. All encoders used their default parameter settings to level the grounds. The number of epochs was maintained at two so that a maximum training time of ten days was imposed. In practice, no substantial increase was experienced when going beyond one epoch, but we opted to give all transformers sufficient time to converge. The maximum sequence length was set to 512, and sliding windows considered a 0.95 stride. The batch size was set to

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

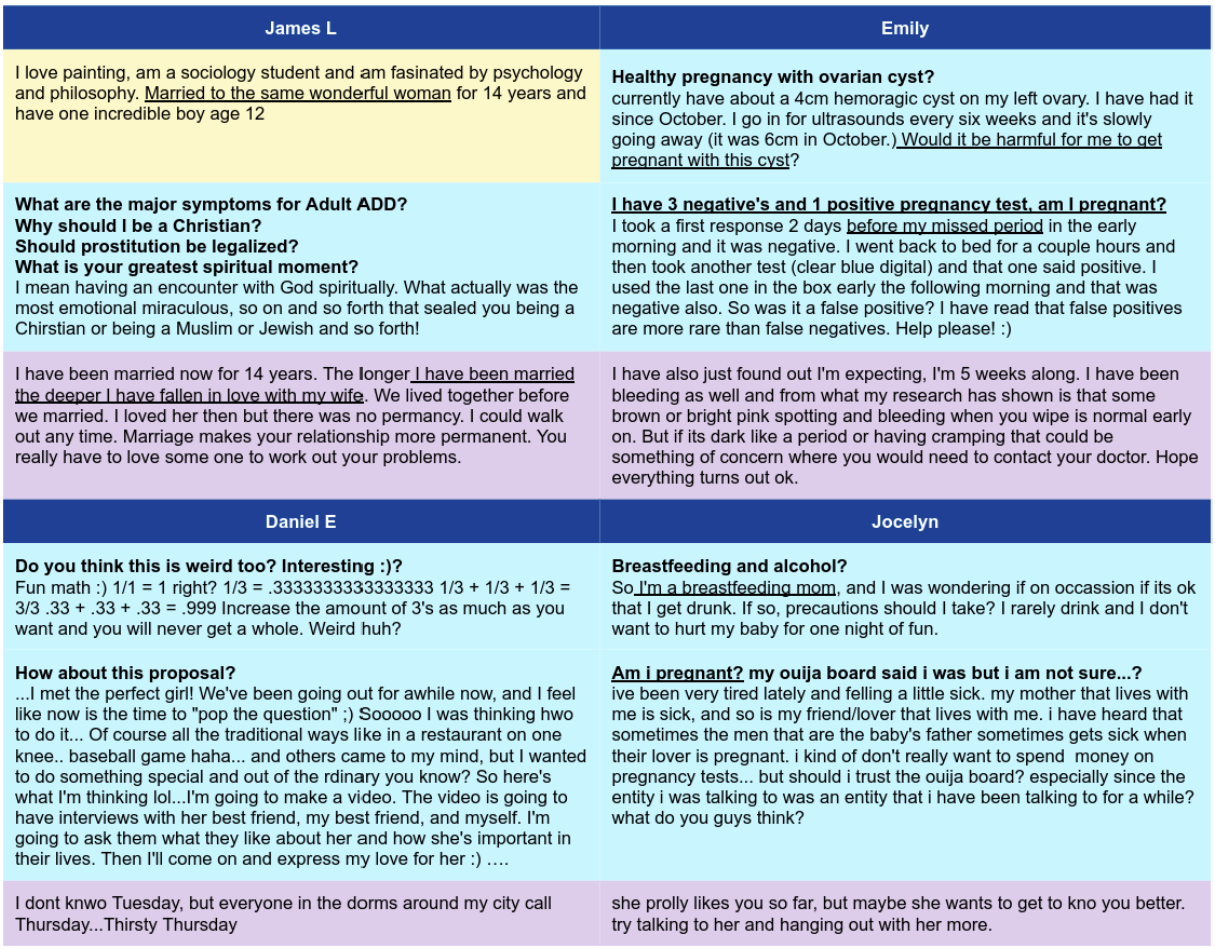

FIGURE 1. Illustrative excerpts from four automatically annotated profiles within our study corpus. On the left, two community members identified as masculine (i.e., "James L" and "Daniel E"), whereas on the right two individuals recognized as feminine (i.e., "Emily" and "Jocelyn"). Light blue and light yellow indicate self-descriptions and questions, respectively. Light purple signals answers, and in bold, question titles. Most relevant cues to guess genders are underlined.

ensure that the GPU[3] memory was fully utilized. We used a batch size of eight except for large models (xlnet, xlm-roberta, deberta, etc.), where a batch size of 128 was required to allow convergence.

V. EXPERIMENTS

To construct our dataset, we used 657,805 member profiles distilled from Yahoo! Answers, a corpus that was previously utilized in [48] and [49] (see figure 1) for age analysis. For annotation purposes (i.e. male or female), we capitalized on a series of heuristics [28]. First, we checked for user aliases contained in any of a group of seven publicly available gender by name collections including WGND[4] and Howarder.[5] As a means of improving the matching, nicknames were lowercased, trimmed at their first space, hyphen, at, underscore or dot. Accordingly, ASCII characters outside the range from 97 to 122 were removed. In the event of no alignments, the end of the alias was systematically trimmed one character at a time until a match was found, or its length was five characters. The final decision was made by counting the overall frequency of each gender.

Second, lowercased n-grams[6] were extracted from all textual content by substituting numbers with a placeholder and ranked in conformity with their entropy afterwards. After a manual inspection of low-ranked elements, almost 1,500 gender indicative phrases were compiled, and later used to revise the previous frequency counts, thus a final label was assigned to 548,375 (83.36%) out of the 657,805 fellows in the corpus (see statistics on table 2). The overall distribution was as follows: 343,661 (62.67%) are women and 204,714 (37.33%) men. The dataset was divided into 329,025 training, 109,675 evaluation and 109,675 testing instances using a random stratified sampling strategy, maintaining on each set the proportions of women and men. From the 109,675 instances in the test set, 68,676 (62,6%) were women and 40,999 men (37.4%). Every piece of text used by the heuristics was

[3] Nvidia DGX A100 2 × 40gb GPUs.
[4] github.com/lizhi1104/nlp_data
[5] data.world/howarder/gender-by-name
[6] stanfordnlp.github.io/CoreNLP/

TABLE 1. Pre-trained models used in the experiments.

Model	Size	Cased	Trained on
albert-base-v1	12M	No	BookCorpus, English Wikipedia
bert-base-cased	110M	Yes	BookCorpus, English Wikipedia
deberta-base	134M	Yes	BookCorpus, English Wikipedia, OpenWebText, STORIES
distilbert-base-cased	66M	Yes	BookCorpus, English Wikipedia
distilroberta-base	82M	Yes	OpenWebText
electra-base	110M	Yes	BookCorpus, English Wikipedia
fnet-base	83M	No	C4
longformer-base-4096	125M	Yes	BookCorpus, English Wikipedia, RealNews, STORIES
roberta-base	125M	Yes	BookCorpus, English Wikipedia, OpenWebText, STORIES, CC-News
mobilebert-uncased	25.3M	No	BookCorpus, English Wikipedia
xlm-roberta-base	270M	Yes	CommonCrawl (100 languages)
xlnet-base-cased	110M	Yes	BookCorpus, English Wikipedia

TABLE 2. General description of the corpus.

Overview	
no. questions	6,690,999
no. questions (title only)	545,634
no. answers	24,064,525
no. best answers	4,286,725
no. self-descriptions	49,507
Amount of tokens	
titles	71,334,205
bodies	633,052,979
answers	1,495,060,174
self-descriptions	2,074,167
Number of sentences	
titles	6,713,913
bodies	15,764,668
answers	56,419,925
self-descriptions	111,904
Unique terms (dictionary size)	
titles	698,998
bodies	2,480,184
answers	6,195,619
self-descriptions	84,220

removed from the respective profile of the user. Note that held-out evaluations were carried out in all our experiments by preserving these splits unchanged. The following abbreviations denote different empirical settings, designed to allow the identification of the individual contribution of each piece of information to classifier performance:

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

Finally, we took advantage of the test samples exclusively to make an unbiased assessment of the final model fit on the training/evaluation instances.

To answer **RQ1 & RQ2**, we fine-tuned several pre-trained transformers 1 using each of the four datasets described earlier, namely TBAD, TBA, TB and T. Our outcomes point towards MobileBert and DeBERTa as the best options for this task, since both achieved a superior performance regardless of the metric and configuration (see tables 3, 4 and 5). Adding self-descriptions does not improved significantly the quality of classification of MobileBert and DeBERTa. It is worth stressing here that only about 7% of the members within this corpus provide a self-description. We conjecture that this low proportion of examples is, in part, one of the reasons for the observed effect on the performance. Overall, these two fine-tuned encoders achieved an Accuracy, F1-Score, and AUC greater than 86.4%, 0.894, 0.921, respectively.

The worst performance can be, most of the time, attributed to one out of the three following models: XLNet, BERT and ELECTRA. In particular, XLNet obtained the lowest scores when trained on question titles (73.82%) and on TBAD (74.15%). Our results suggest that one reason to this may be their sensitivity to the distinct input signals. Interestingly, these three alternatives are more competitive under a TBA configuration, but on the other hand, their performance significantly drop when considering self-descriptions or when discarding answers. For the most part, the gap between the best and the worst systems is the narrowest when fine-tuning using full questions and answers (approximately 6.6% accuracy). In contrast, training solely on full questions brings about the widest gap (approximately 14.92% accuracy).

Furthermore, Table 3 indicates that the average accuracy was 0.8051, with a maximum of 86.66% (DeBERTa) and a minimum of 69.93% (ELECTRA). Our dataset was imbalanced because of that, in addition to accuracy, we reported the f1 metric (Table 5) and AUC Score (Table 4), because those metrics are appropriate to compare classifiers in presence of imbalanced data [50]. Table 4 shows that the average AUC score was 0.8562, ranging from 0.6745 (XLNet) to 0.9069 (MobileBERT). Overall, these results show that it is possible to detect gender differences across community peers using textual interactions within the cQA site. They also revealed that models designed for efficiency and that are case insensitive such as MobileBert obtain the best average results (AUC: 0.9069). To be more specific, this cost-efficiency was observed when juxtaposing the classification rates accomplished using the following settings: DeBERTa (TBA) with an AUC value of 0.9247, closely

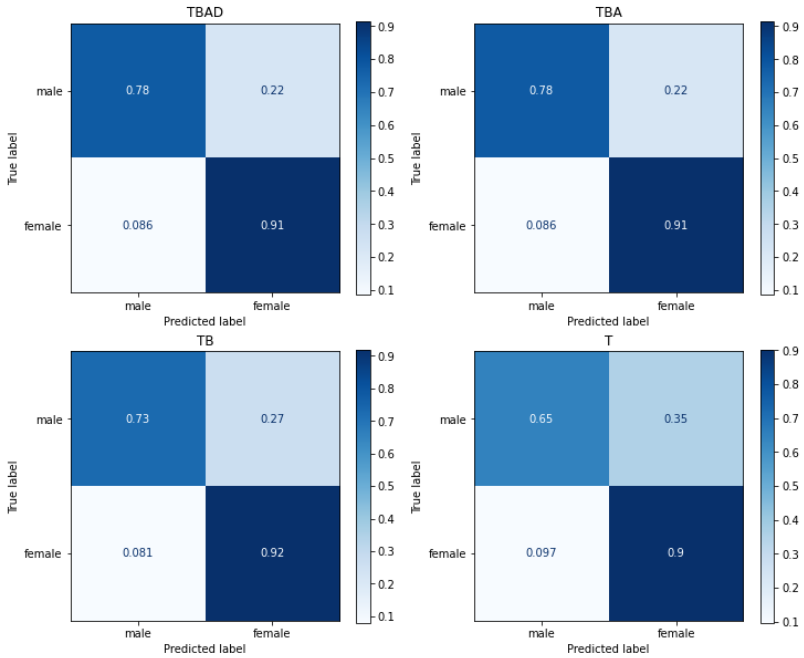

FIGURE 2. Confusion Matrices (MobileBERT).

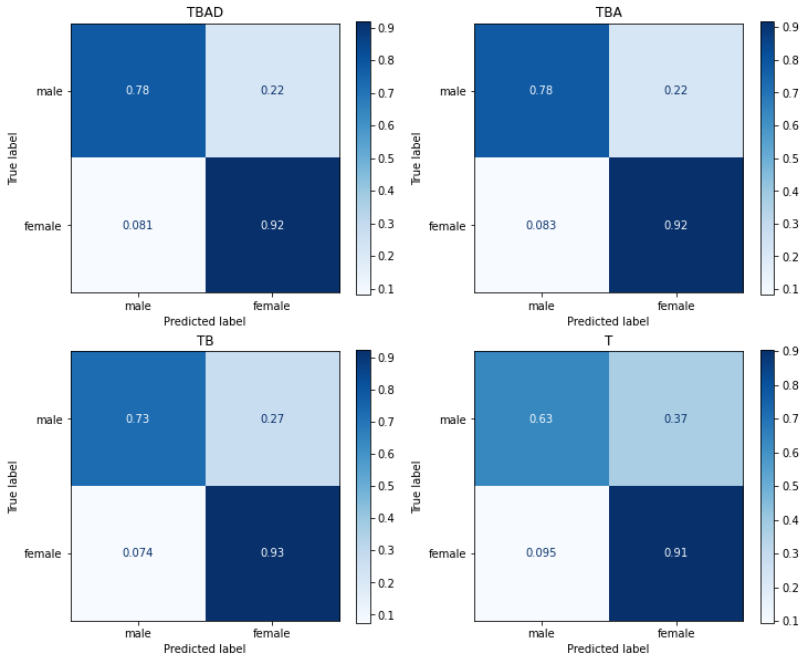

FIGURE 3. Confusion Matrices (DeBERTa).

followed by MobileBERT (TBAD) and (TBA) with AUC scores of 0.9223 and 0.9217, respectively.

For error analysis, figures 2 and 3 show confusion matrices in the test set for MobileBERT and DeBERTa (i.e., TBAD, TBA, TB, T). MobileBERT TBAD classified correctly a 91,4% of feminine profiles (62,773 of 68,676), while the highest number of masculine samples correctly classified was achieved by TBAD and TBA models (78%, 31.979 of 40,999). To further analyze the effect of dataset imbalance on model learning and classification performance, we trained MobileBERT uncased on two balanced versions of the TBAD dataset, with a 1:1 feminine/masculine proportion, created

TABLE 3. Results in terms of accuracy obtained by each transformer under each of the four different pre-defined settings (test set).

Model	TBAD	TBA	TB	T	Max.	Avrg.	Min.	σ
albert-base-v1	0.7720	0.8222	0.8273	0.7852	0.8273	0.8017	0.7720	0.0236
bert-base-cased	0.8136	0.8006	0.7465	0.7537	0.8136	0.7786	0.7465	0.0290
deberta-base	0.8621	**0.8666**	0.8347	**0.8117**	**0.8666**	0.8438	0.8117	0.0222
distilbert-base-cased	0.8192	0.8074	0.8338	0.7960	0.8338	0.8141	0.7960	**0.0140**
distilroberta-base	0.8366	0.8298	0.8438	0.7998	0.8438	0.8275	0.7998	0.0167
electra-base	0.8096	0.8143	0.6993	0.7571	0.8143	0.7701	**0.6993**	0.0466
fnet-base	0.8427	0.8415	0.8252	0.7918	0.8427	0.8253	0.7918	0.0205
longformer-base-4096	0.7890	0.8040	0.7978	0.7735	0.8040	0.7911	0.7735	0.0114
roberta-base	0.7897	0.8344	0.7404	0.7683	0.8344	0.7832	0.7404	0.0343
mobilebert-uncased	**0.8627**	0.8642	**0.8485**	0.8075	0.8642	**0.8457**	0.8075	0.0229
xlm-roberta-base	0.8602	0.8070	0.7790	0.7404	0.8602	0.7966	0.7404	0.0437
xlnet-base-cased	0.7415	0.8596	0.7924	0.7382	0.8596	0.7829	0.7382	0.0492
Max.	0.8627	0.8666	0.8485	0.8117	0.8666	0.8457	0.8117	0.0492
Avrg.	0.8166	0.8293	0.7974	0.7769	0.8387	0.8051	0.7681	0.0279
Min.	0.7415	0.8006	0.6993	0.7382	0.8040	0.7701	0.6993	0.0114
σ	0.0369	0.0231	0.0456	0.0245	0.0204	0.0247	0.0335	0.0123

TABLE 4. Outcomes in terms of AUC (Area Under the Curve) accomplished by each combination of encoder and input (test set).

Model	TBAD	TBA	TB	T	Max.	Avrg.	Min.	σ
albert-base-v1	0.8273	0.8871	0.8905	0.8461	0.8905	0.8627	0.8273	0.0269
bert-base-cased	0.8750	0.8543	0.7986	0.7736	0.8750	0.8254	0.7736	0.0409
deberta-base	0.9205	**0.9247**	0.9001	**0.8759**	**0.9247**	0.9053	0.8759	0.0193
distilbert-base-cased	0.8812	0.8661	0.8949	0.8575	0.8949	0.8749	0.8575	**0.0143**
distilroberta-base	0.9029	0.8673	0.9017	0.8619	0.9029	0.8834	0.8619	0.0189
electra-base	0.8462	0.8383	0.8238	0.7896	0.8462	0.8245	0.7896	0.0217
fnet-base	0.9060	0.9033	0.8897	0.8541	0.9060	0.8883	0.8541	0.0207
longformer-base-4096	0.7657	0.8535	0.8442	0.8274	0.8535	0.8227	0.7657	0.0342
roberta-base	0.8170	0.8705	0.8524	0.8186	0.8705	0.8396	0.8170	0.0228
mobilebert-uncased	**0.9223**	0.9217	**0.9116**	0.8720	0.9223	**0.9069**	0.8720	0.0206
xlm-roberta-base	0.9197	0.8786	0.8023	0.7695	0.9197	0.8425	0.7695	0.0596
xlnet-base-cased	0.7293	0.9175	0.8715	0.6745	0.9175	0.7982	**0.6745**	0.0996
Max.	0.9223	0.9247	0.9116	0.8759	0.9247	0.9069	0.8759	0.0996
Avrg.	0.8594	0.8819	0.8651	0.8184	0.8936	0.8562	0.8115	0.0333
Min.	0.7293	0.8383	0.7986	0.6745	0.8462	0.7982	0.6745	0.0143
σ	0.0611	0.0277	0.0383	0.0560	0.0259	0.0342	0.0571	0.0233

TABLE 5. F1 scores for each transformer vs. each pre-defined configuration (test set).

Model	TBAD	TBA	TB	T	Max.	Avrg.	Min.	σ
albert-base-v1	0.8341	0.8625	0.8685	0.8391	0.8685	0.8511	0.8341	0.0147
bert-base-cased	0.8569	0.8497	0.7848	0.8159	0.8569	0.8268	0.7848	0.0288
deberta-base	0.8927	**0.8960**	0.8741	**0.8580**	**0.8960**	0.8802	0.8580	0.0153
distilbert-base-cased	0.8609	0.8538	0.8734	0.8461	0.8734	0.8585	0.8461	**0.0100**
distilroberta-base	0.8693	0.8710	0.8804	0.8500	0.8804	0.8677	0.8500	0.0110
electra-base	0.8533	0.8588	0.7182	0.8233	0.8588	0.8134	**0.7182**	0.0566
fnet-base	0.8751	0.8771	0.8699	0.8415	0.8771	0.8659	0.8415	0.0143
longformer-base-4096	0.8424	0.8510	0.8504	0.8319	0.8510	0.8439	0.8319	0.0077
roberta-base	0.8495	0.8745	0.8251	0.8318	0.8745	0.8452	0.8251	0.0191
mobilebert-uncased	**0.8929**	0.8940	**0.8837**	0.8546	0.8940	**0.8813**	0.8546	0.0159
xlm-roberta-base	0.8912	0.8407	0.8335	0.8173	0.8912	0.8457	0.8173	0.0276
xlnet-base-cased	0.7967	0.8898	0.8512	0.8172	0.8898	0.8388	0.7967	0.0353
Max.	0.8929	0.8960	0.8837	0.8580	0.8960	0.8813	0.8580	0.0566
Avrg.	0.8596	0.8683	0.8428	0.8356	0.8760	0.8515	0.8215	0.0214
Min.	0.7967	0.8407	0.7182	0.8159	0.8510	0.8134	0.7182	0.0077
σ	0.0267	0.0177	0.0464	0.0143	0.0145	0.0195	0.0378	0.0133

using random oversampling and random undersampling [51]. Both models achieved and AUC score of 0.92 and an accuracy of 0.85 on the balanced test set, similar to MobileBERT trained on the imbalanced dataset.

TABLE 6. Case analysis for DistilBERT, DistilRoBERTa, XLM-RoBERTa and DeBERTa. The star denotes the use of CoreNLP case corrector when fine-tuning.

	TBAD			TBA			TB			T		
Model	F1	Acc.	AUC	F1	Acc.	AUC	F1	Acc.	AUC	F1	Acc.	AUC
distilbert-base-uncased	0.880	0.845	0.906	0.877	0.843	0.906	0.866	0.823	0.886	0.846	0.797	0.858
distilbert-base-cased	0.861	0.819	0.881	0.854	0.807	0.866	0.873	0.834	0.895	0.846	0.796	0.857
distilbert-base-cased*	0.869	0.831	0.892	0.866	0.828	0.892	0.864	0.823	0.887	0.835	0.780	0.839
distilroberta-base	0.869	0.837	0.903	0.871	0.830	0.867	0.880	0.844	0.902	0.850	0.800	0.862
distilroberta-base*	0.885	0.851	0.911	0.881	0.846	0.908	0.873	0.834	0.897	0.845	0.793	0.853
xlm-roberta-base	0.891	0.860	0.920	0.841	0.807	0.879	0.834	0.779	0.802	0.817	0.740	0.769
xlm-roberta-base*	0.892	0.860	0.921	0.892	0.860	0.920	0.863	0.820	0.886	0.853	0.805	0.869
deberta-base	0.893	0.862	0.921	0.896	0.867	0.925	0.874	0.835	0.900	0.858	0.812	0.876
deberta-base*	0.897	0.868	0.926	0.896	0.866	0.925	0.887	0.852	0.915	0.852	0.803	0.868

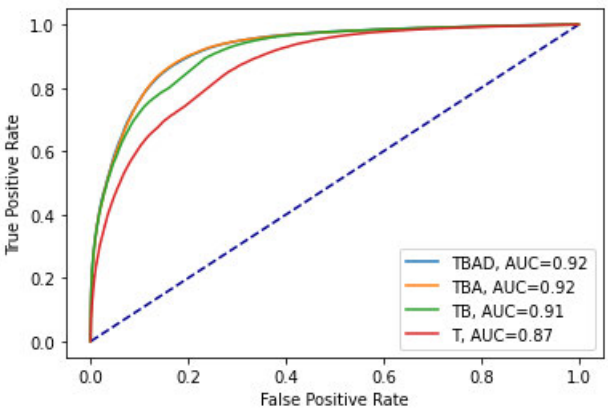

FIGURE 4. ROC Curves (MobileBERT).

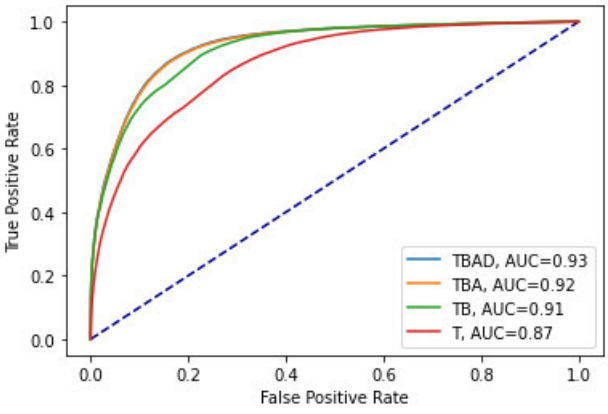

FIGURE 5. ROC Curves (DeBERTa).

In summary, testing assorted fine-tuned transformers is pertinent to gender recognition across cQA platforms, as our experiments revealed a high variability in the classification rates among distinct encoders under the same input signals. We deem this to be a result of their sharply different specialized designs with different vocabulary sizes. Owing to their architectural differences, there is no rule for determining the right transformer and configuration for a particular downstream task.

To answer **RQ3**, the data in Tables 3, 4 and 5 show that the best results were obtained using a combination of questions and answers (TBA, TBAD), but the best results on average were obtained using only question titles, question bodies and answers on TBA models (AUC: 0.8819, accuracy: 0.8293, F1: 0.8683). These models achieve a balance between accuracy, precision and recall. Models that included question bodies performed 5% better than models trained on question titles only (AUC: 0.8651 vs AUC: 0.8184). With the inclusion of answers, the models exhibited an average increase of 2% in performance (AUC: 0.8819). The inclusion of profile descriptions led to a decrease in the average performance (AUC: 0.8594), although some models exhibited a small positive effect. Considering that only 7% of the profiles had descriptions, these results suggest that the inclusion of profile descriptions may be omitted without significantly affecting performance.

Regarding **RQ4**, uncased models (mobilebert-uncased, fnet-base) tend to perform better in our cQA dataset, which suggests that the writing of questions and answers in the cQA site is different from that in clean corpora, where most models were originally trained (e.g., BookCorpus and English Wikipedia), and perhaps less formal. The use of uncased models appears to mitigate some of the differences in writing between the datasets. Table 6 shows a comparison of the results obtained by the cased and uncased versions of the models with higher average performance, trained on the raw dataset and a corrected case dataset (true case). Distilbert-base-uncased performed slightly better than its cased counterparts on the raw dataset. When the case is corrected, the distilbert-base-cased model performs better than when trained on the raw dataset but is still below the performance of distilbert-base-uncased. This pattern also occurs on distilroberta and xlm-roberta, the models performed better when trained on the dataset with the corrected case.

The results obtained using the deberta-base cased model, require further analysis. Deberta results may be explained because is trained on OpenWebText and STORIES. OpenWebText is a corpus generated from reddit, a social media platform, where users can add their own content, and other users can qualify that content, probably leading to writing more similar to a cQA site. This relative similarity in content

may lead to deberta to produce a language representation that might be best suited to learning how to represent the cQA content. The Deberta-base model does not perform significantly better when trained on the true case dataset. The AUC of the TBA model did not improve in relation to the raw dataset, and the TBAD model performance improved from 0.921 to 0.926. The Deberta-base model trained on the true case dataset achieved the best score of all trained models, with an average AUC of 0.9085, slightly above that of the mobilebert-base model trained on the raw dataset, with an AUC of 0.9069. Nevertheless, the complexity of Deberta is five times larger than that of MobileBert.

Figure 3 shows confusion matrices for deberta-base trained on the true case dataset. Their error rates are very similar to the ones of MobileBERT, only the TB model shows a 1% percent increase on their female detection capability. Regarding ROC Curves, Figures 5 and 4 shows that the performance of both models is indistinguishable.

Based on these results, we selected mobilebert-uncased TBA as the model with the best balance between performance and complexity.

A. VISUALIZATION AND EXPLANATION OF MODEL BEHAVIOR

Explainable AI (XAI) [52], [53], [54] and explainability of transformers [55], [56], [57], [58], [59], [60] are active areas of research. There are multiple approaches to construct explanations for transformer models, many of them relying on the use of attention weights. To understand the classifications provided by our models, we analyzed the mobileBERT model (T and TB versions) to explain the classification of samples for users Daniel and Emily shown in figure 1, that were not included in the samples used for training.

We selected the sample for Daniel because was mislabeled by the T model but correctly classified by the TB model, allowing us to analyze the reason behind the improvement. The sample for Emily was correctly classified by both models and allow us to analyze if the reason for classification changed between the T and TB models. For model understanding we used attention visualization [55], [57], [58] and attribution [56] using the python package transformers-interpret.[7]

Figures 6 and 7 show the attention weights in the first layer of mobileBERT T model for head 1 (Top) and head 2 (Bottom). On both figures we see two common attention patterns [58], head 1 has a heterogeneous pattern and head 2 a diagonal pattern. The heterogeneous pattern of head 1 suggest that the model learned semantic relations, like the relation between the word 'is' with words 'weird' and 'interesting', on figure 6. The diagonal pattern appears when attention is between previous, current and next words, like the relation between 'am' and 'pregnant' on head 2 depicted in figure 7.

The sample for the masculine user was incorrectly classified by the mobileBERT T model and correctly classified

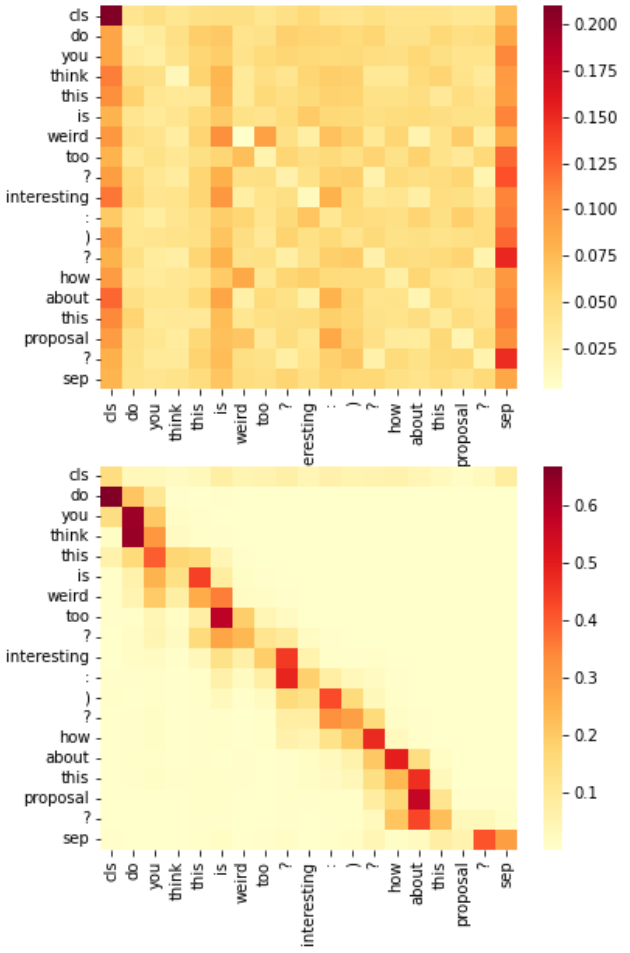

FIGURE 6. Attention for the sample of user Daniel.

by the TB model. To explain the difference, Figure 8 shows the attribution scores [56] of the T model (Top) and TB model (Bottom). Highlighted in green are the words that influenced the conclusion reached by the model, while in red are the words supporting the other option, discarded by the model. The T model classified the masculine sample as feminine based mainly in the use of the word interesting and the emoticon :) (attribution 0.16 and 0.32, respectively). This behaviour is concordant with previous analyses [30] that found and association between female interventions and positive sentiments. The TB model did the correct classification as masculine, influenced by words met, girl and baseball (attributions 0.53, 0.39 and 0.23). For the feminine sample, the words healthy, pregnancy and pregnant (attributions 0.28, 0.28 and 0.47) influenced the classification on both models, while the words period and couple (attributions 0.20 and 0.34) complemented the decision on the TB model. These behaviors are more related with the topic discussed by the users, and the use of certain words when talking about something (meeting a woman, pregnancy) or when referring to self (pregnant). In both cases, the results suggest that models learned a relationship between some topics and gender, based

[7] https://github.com/cdpierse/transformers-interpret.git

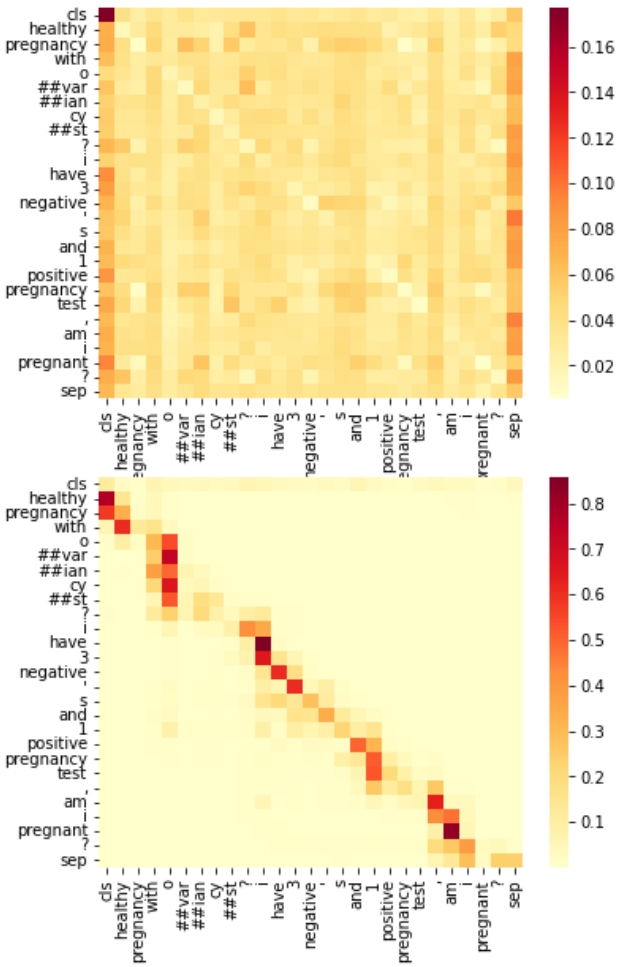

FIGURE 7. Attention for the sample of user Emily.

[CLS] do you think this is weird too ? interesting :) ? how about this proposal ? [SEP]

[CLS] do you think this is weird too ? interesting :) ? fun math :) 1 / 1 right ? 1 / 3 = . 333 ##33 ##33 ##33 ##33 ##33 ##33 ##3 1 / 3 + 1 / 3 + 1 / 3 = 3 / 3 . 33 + . 33 + . 33 = . 999 increase the amount of 3 ' s as much as you want and you will never get a whole . weird huh ? how about this proposal ? . . . i met the perfect girl ! we ' ve been going out for awhile now , and i feel like now is the time to " pop the question " ;) soo ##oo ##o i was thinking h ##wo to do it . . . of course all the traditional ways like in a restaurant on one knee . . baseball game ha ##ha . . . and others came to my mind , but i wanted to do something special and out of the rd ##ina ##ry you know ? so here ' s what i ' m thinking lo ##l . . . i ' m going to make a video . the video is going to have interviews with her best friend , my best friend , and myself . i ' m going to ask them what they like about her and how she ' s important in their lives . then i ' ll come on and express my love for her :) . . . [SEP]

FIGURE 8. Attribution for sample of user Daniel.

[CLS] healthy pregnancy with o ##var ##ian cy ##st ? i have 3 negative ' s and 1 positive pregnancy test , am i pregnant ? [SEP]

[CLS] healthy pregnancy with o ##var ##ian cy ##st ? currently have about a 4 ##cm hem ##ora ##gic cy ##st on my left o ##vary . i have had it since october . i go in for ultrasound ##s every six weeks and it ' s slowly going away (it was 6 ##cm in october .) would it be harmful for me to get pregnant with this cy ##st ? i have 3 negative ' s and 1 positive pre ##gan ##ancy test , am i pregnant ? i took a first response 2 days before my missed period in the early morning and it was negative . i went back to bed for a couple hours and then took another test (clear blue digital) and that one said positive . i used the last one in the box early the following morning and that was negative also . so was it a false positive ? i have read that false positive ##s are more rare than false negative ##s . help please ! :) [SEP]

FIGURE 9. Attribution for sample of user Emily.

on the information inside the corpus used for training. Previous works had found topic and intent differences between masculine and feminine participants in cQA sites [31].

B. CAVEATS

Genders were assigned according to how members identified themselves on the website. As a rule of thumb, we manually assessed 100 randomly chosen labelled profiles, and obtained an error rate of 10%. Aside from errors attributed to the intrinsic shallow nature of our heuristics, some individuals run fake profiles. Our preliminary manual inspection did not find that other sexual orientations made up a substantial share of this dataset. Owing to their discretion and/or low participation, it is also difficult to compile a comprehensive list of their typifying names and phrases.

VI. LIMITATIONS AND FUTURE RESEARCH

Apart from the aforementioned considerations, there are some additional aspects that must be weighed carefully. First, self-descriptions suffer severely from data-sparseness, namely a low percentage of the members (7%) provides this short biography on Yahoo! Answers. Intuitively, one can expect a great probability of finding pieces of information conveying demographics across this sort of text. Therefore, its real impact should be quantified by studying platforms, where their users are more likely to describe themselves. Here, one could think on services such as Reddit and Stack Exchange. In the same spirit, different ways of integrating this class of input into a, probably joint, model can be further explored in future works.

Second, if significant computational power and massive cQA collections are accessible, one could think about pre-training frontier transformers on user-generated cQA texts. Doing this poses several exciting challenges, for instance to clean or not to clean the corpus? When these resources are inaccessible, the transfer of knowledge can still be improved by means of resolving community jargon, spellings, aliases, entities and acronyms. Additionally, we conjecture that pre-training title-only models will be beneficial, but this will need special adjustments, since the

grammar exhibited in question titles is sharply different to what we can find across question bodies and answers.

Lastly, we also envision that exploiting multi-lingual architectures and texts written in different languages can help to enhance the gender detection rate, especially across users linked to very few questions and answers posted in English. On top of that, multilingualism might assist in dealing with the data-sparseness in self-descriptions. In the case of Yahoo! Answers, extra biographies can be harvested from some Spanish speaking members. Recall that our study focused its attention only on textual content in English, which was singled out via a language detector.

VII. CONCLUSION

Regarding **RQ1**, we concluded that it is possible to infer the gender of a community peer on a cQA site from their interactions with the site. Better results were obtained by models using full questions (title and body) combined with the answers provided by the person. Uncased models (i.e., Mobilebert and FNET) and models trained on varied datasets like Deberta, performed better than models trained on clean corpora such as BookCorpus or English Wikipedia, showing that the use of any pre-trained model does not lead to the same classification results (**RQ2**, **RQ3**). Another important conclusion is that the addition of more information does not always lead to better results, because some TBA models performed better than their TBAD counterparts. The differences in results may be explained by differences in writing across datasets (RQ4), because the correction of the case of the words improved the results in cased models. This affirmation could be further investigated on future works, by training the models with an updated dataset where misspelled words are corrected to ease the transfer learning from the clean corpus.

Model selection appears to be an important issue in the context of natural language understanding applied to cQA sites. We summarize our findings in the following guidelines for model reuse:

- To improve transfer learning, select a model trained on a dataset with the closest similarity in writing (formal, informal) to the dataset used for the downstream task.
- If the selected model is cased, preprocess the dataset to correct case before training.
- When using a model pre-trained on a clean corpus, consider fine-tuning the model using a dataset where misspelled words are corrected.

We conclude that gender recognition based on writing may be helpful in profiling users in cQA sites, and as a tool to design interventions to promote equal engagement and participation in online communities.

REFERENCES

[1] A. Figueroa, "Male or female: What traits characterize questions prompted by each gender in community question answering?" *Exp. Syst. Appl.*, vol. 90, pp. 405–413, Dec. 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417417305845

[2] T. B. Brown et al., "Language models are few-shot learners," 2020, *arXiv:2005.14165*.

[3] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," 2020, *arXiv:2005.12872*.

[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. 31st Int. Conf. Neural Inf. Process. Syst.* Red Hook, NY, USA: Curran Associates, 2017, pp. 6000–6010.

[5] T. Lin, Y. Wang, X. Liu, and X. Qiu, "A survey of transformers," 2021, *arXiv:2106.04554*.

[6] C. Sun, X. Qiu, Y. Xu, and X. Huang, "How to fine-tune BERT for text classification?" 2019, *arXiv:1905.05583*.

[7] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *J. Mach. Learn. Res.*, vol. 21, no. 140, pp. 1–67, 2020. [Online]. Available: http://jmlr.org/papers/v21/20-074.html

[8] A. Radford and K. Narasimhan, "Improving language understanding by generative pre-training," Tech. Rep., 2018. [Online]. Available: https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf

[9] 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.*, vol. 1, Jun. 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423

[10] X. Qiu, T. Sun, Y. Xu, Y. Shao, N. Dai, and X. Huang, "Pre-trained models for natural language processing: A survey," *Sci. China Technol. Sci.*, vol. 63, pp. 872–1897, Sep. 2020.

[11] T. Lin, Y. Wang, X. Liu, and X. Qiu, "A survey of transformers," *AI Open*, vol. 3, pp. 111–132, 2022.

[12] G. Brauwers and F. Frasincar, "A general survey on attention mechanisms in deep learning," *IEEE Trans. Knowl. Data Eng.*, early access, Nov. 9, 2021, doi: 10.1109/TKDE.2021.3126456.

[13] Y. Tay, M. Dehghani, D. Bahri, and D. Metzler, "Efficient transformers: A survey," *ACM Comput. Surveys*, vol. 55, no. 6, pp. 1–28, Jul. 2023.

[14] X. Han et al., "Pre-trained models: Past, present and future," 2021, *arXiv:2106.07139*.

[15] Y. Lin, Y. C. Tan, and R. Frank, "Open sesame: Getting inside BERT's linguistic knowledge," 2019, *arXiv:1906.01698*.

[16] G. Jawahar, B. Sagot, and D. Seddah, "What does BERT learn about the structure of language?" in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 3651–3657. [Online]. Available: https://aclanthology.org/P19-1356

[17] A. Rogers, O. Kovaleva, and A. Rumshisky, "A primer in bertology: What we know about how BERT works," *Trans. Assoc. Comput. Linguistics*, vol. 8, pp. 842–866, Dec. 2020.

[18] I. Tenney, D. Das, and E. Pavlick, "BERT rediscovers the classical NLP pipeline," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 4593–4601. [Online]. Available: https://aclanthology.org/P19-1452

[19] D. Jin, Z. Jin, J. T. Zhou, and P. Szolovits, "Is bert really robust? A strong baseline for natural language attack on text classification and entailment," in *Proc. AAAI Conf. Artif. Intell.*, Apr. 2020, vol. 34, no. 5, pp. 8018–8025. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/6311

[20] A. Talmor, Y. Elazar, Y. Goldberg, and J. Berant, "OLMpics—On what language model pre-training captures," 2019, *arXiv:1912.13283*.

[21] K. Richardson, H. Hu, L. Moss, and A. Sabharwal, "Probing natural language inference models through semantic fragments," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 5, pp. 8713–8721, Apr. 2020. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/6397

[22] J. Wallat, J. Singh, and A. Anand, "BERTnesia: Investigating the capture and forgetting of knowledge in BERT," 2021, *arXiv:2106.02902*.

[23] A. Ahmad, C. Feng, S. Ge, and A. Yousif, "A survey on mining stack overflow: Question and answering (q&a) community," *Data Technol. Appl.*, vol. 52, 2, pp. 190–247, 2018.

[24] A. Bouziane, D. Bouchiha, N. Doumi, and M. Malki, "Question answering systems: Survey and trends," *Proc. Comput. Sci.*, vol. 73, no. 73, pp. 366–375, 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1877050915034663

[25] I. Srba and M. Bielikova, "A comprehensive survey and classification of approaches for community question answering," *ACM Trans. Web*, vol. 10, no. 3, pp. 1–63, Aug. 2016, doi: 10.1145/2934687.

[26] J. M. Jose and J. Thomas, "Finding best answer in community question answering sites: A review," in *Proc. Int. Conf. Circuits Syst. Digit. Enterprise Technol. (ICCSDET)*, Dec. 2018, pp. 1–5. [Online]. Available: https://ieeexplore.ieee.org/document/8821219

[27] P. K. Roy, S. Saumya, J. P. Singh, S. Banerjee, and A. Gutub, "Analysis of community question-answering issues via machine learning and deep learning: State-of-the-art review," *CAAI Trans. Intell. Technol.*, pp. 1–23, May 2022. [Online]. Available: https://ietresearch.onlinelibrary.wiley.com/doi/abs/10.1049/cit2.12081, doi: 10.1049/cit2.12081.

[28] B. Lin and A. Serebrenik, "Recognizing gender of stack overflow users," in *Proc. 13th Int. Conf. Mining Softw. Repositories*. New York, NY, USA: Association for Computing Machinery, May 2016, pp. 425–429, doi: 10.1145/2901739.2901777.

[29] B. Peralta, A. Figueroa, O. Nicolis, and A. Trewhela, "Gender identification from community question answering avatars," *IEEE Access*, vol. 9, pp. 156701–156716, 2021, doi: 10.1109/ACCESS.2021.3130078.

[30] 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 (WSDM)*. New York, NY, USA: ACM, 2012, pp. 633–642, doi: 10.1145/2124295.2124371.

[31] P. M. J. Dubois, M. Maftouni, P. K. Chilana, J. McGrenere, and A. Bunt, "Gender differences in graphic design Q&AS: How community and site characteristics contribute to gender gaps in answering questions," in *Proc. ACM Hum.-Comput. Interact.*, vol. 4, no. CSCW2, 2020, pp. 1–26.

[32] D. Ford, A. Harkins, and C. Parnin, "Someone like me: How does peer parity influence participation of women on stack overflow?" in *Proc. IEEE Symp. Vis. Lang. Human-Centric Comput. (VL/HCC)*, Oct. 2017, pp. 239–243.

[33] A. May, J. Wachs, and A. Hannák, "Gender differences in participation and reward on stack overflow," *Empirical Softw. Eng.*, vol. 24, no. 4, pp. 1997–2019, Aug. 2019.

[34] Y. Wang, "Understanding the reputation differences between women and men on stack overflow," in *Proc. 25th Asia–Pacific Softw. Eng. Conf. (APSEC)*, Dec. 2018, pp. 436–444.

[35] G. Blanco, R. Pérez-López, F. Fdez-Riverola, and A. M. G. Lourenço, "Understanding the social evolution of the Java community in stack overflow: A 10-year study of developer interactions," *Future Gener. Comput. Syst.*, vol. 105, pp. 446–454, Apr. 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167739X19311884

[36] S. J. Brooke, "Trouble in programmer's paradise: Gender-biases in sharing and recognising technical knowledge on stack overflow," *Inf., Commun. Soc.*, vol. 24, no. 14, pp. 2091–2112, Oct. 2021.

[37] W. L. Taylor, "Cloze procedure: A new tool for measuring readability," *Journalism Bull.*, vol. 30, no. 4, pp. 415–433, 1953.

[38] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "ALBERT: A lite BERT for self-supervised learning of language representations," in *Proc. 8th Int. Conf. Learn. Represent. (ICLR)*, Addis Ababa, Ethiopia, Apr. 2020, pp. 1–17. [Online]. Available: https://openreview.net/forum?id=H1eA7AEtvS

[39] P. He, X. Liu, J. Gao, and W. Chen, "DeBERTa: Decoding-enhanced BERT with disentangled attention," 2020, *arXiv:2006.03654*.

[40] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter," 2019, *arXiv:1910.01108*.

[41] K. Clark, M.-T. Luong, Q. V. Le, and C. D. Manning, "ELECTRA: Pre-training text encoders as discriminators rather than generators," in *Proc. ICLR*, 2020, pp. 1–18. [Online]. Available: https://openreview.net/pdf?id=r1xMH1BtvB

[42] J. Lee-Thorp, J. Ainslie, I. Eckstein, and S. Ontanon, "FNet: Mixing tokens with Fourier transforms," 2021, *arXiv:2105.03824*.

[43] I. Beltagy, M. E. Peters, and A. Cohan, "Longformer: The long-document transformer," 2020, *arXiv:2004.05150*.

[44] Z. Sun, H. Yu, X. Song, R. Liu, Y. Yang, and D. Zhou, "MobileBERT: A compact task-agnostic BERT for resource-limited devices," 2020, *arXiv:2004.02984*.

[45] 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*.

[46] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, "XLNet: Generalized autoregressive pretraining for language understanding," in *Advances in Neural Information Processing Systems*, vol. 32, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds. Red Hook, NY, USA: Curran Associates, 2019.

[47] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," 2019, *arXiv:1911.02116*.

[48] 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, 2021. [Online]. Available: https://www.mdpi.com/2078-2489/12/2/48

[49] A. Figueroa and M. Timilsina, "What identifies different age cohorts in yahoo! Answers?" *Knowl.-Based Syst.*, vol. 228, Sep. 2021, Art. no. 107278. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0950705121005402

[50] Y. Sun, A. K. Wong, and M. S. Kamel, "Classification of imbalanced data: A review," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 23, no. 4, pp. 687–719, 2009.

[51] R. Mohammed, J. Rawashdeh, and M. Abdullah, "Machine learning with oversampling and undersampling techniques: Overview study and experimental results," in *Proc. 11th Int. Conf. Inf. Commun. Syst. (ICICS)*, Apr. 2020, pp. 243–248.

[52] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Inf. Fusion*, vol. 58, pp. 82–115, Jun. 2020.

[53] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52138–52160, 2018.

[54] F. Cabitza, A. Campagner, G. Malgieri, C. Natali, D. Schneeberger, K. Stoeger, and A. Holzinger, "Quod erat demonstrandum?—Towards a typology of the concept of explanation for the design of explainable AI," *Exp. Syst. Appl.*, vol. 213, Mar. 2023, Art. no. 118888.

[55] S. Abnar and W. Zuidema, "Quantifying attention flow in transformers," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 4190–4197.

[56] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *Proc. Int. Conf. Mach. Learn. (PMLR)*, 2017, pp. 3319–3328.

[57] J. Vig, "A multiscale visualization of attention in the transformer model," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics, Syst. Demonstrations*, 2019, pp. 37–42.

[58] O. Kovaleva, A. Romanov, A. Rogers, and A. Rumshisky, "Revealing the dark secrets of BERT," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, 2019, pp. 4365–4374.

[59] K. Clark, U. Khandelwal, O. Levy, and C. D. Manning, "What does BERT look at? An analysis of bert's attention," in *Proc. ACL Workshop BlackboxNLP, Analyzing Interpreting Neural Netw. (NLP)*, 2019, pp. 276–286.

[60] A. Coenen, E. Reif, A. Yuan, B. Kim, A. Pearce, F. Viégas, and M. Wattenberg, "Visualizing and measuring the geometry of bert," in *Proc. 33rd Int. Conf. Neural Inf. Process. Syst.*, 2019, pp. 8594–8603.

PABLO SCHWARZENBERG (Member, IEEE) received the Ph.D. degree in computer science from the Pontificia Universidad Católica de Chile. He is currently the Head of the Computer Science Engineering Program, Faculty of Engineering, Universidad Andrés Bello, Santiago, Chile. His research interests include learning analytics and applications of artificial intelligence in education to promote engagement and handle individual differences in massive courses.

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