Language as a fingerprint: Self-supervised learning of user encodings using transformers

Roberta Rocca
Aarhus University
University of Texas at Austin
roberta.rocca@cas.au.dk

Tal Yarkoni
University of Texas at Austin
tyarkoni@utexas.edu

Abstract

The way we talk carries information about who we are. Demographics, personality, clinical conditions, political preferences influence what we speak about and how, suggesting that many individual attributes could be inferred from adequate encodings of linguistic behavior. Conversely, conditioning text representations on author attributes has been shown to improve model performance in many NLP tasks. Previous research on individual differences and language representations has mainly focused on predicting selected attributes from text, or on conditioning text representations on such attributes for author-based contextualization. Here, we present a *self-supervised* approach to learning language-based user encodings using transformers. Using a large corpus of Reddit submissions, we fine-tune DistilBERT on user-based triplet loss. We show that fine-tuned models can pick up on complex linguistic signatures of users, and that they are able to infer rich information about their profiles. Through a series of intrinsic analyses and probing tasks, we provide evidence that fine-tuning enhances models' ability to abstract generalizable user information, which yields performance advantages for user-based downstream tasks. We discuss applications in language-based assessment and contextualized and personalized NLP.

1 Introduction

Language is not simply a means to communicate about events or inner states. What we talk about, with whom, and in what way says something about who we are – our demographics, our personality, our social and political identity. People with different political views are likely to describe political events in radically different ways; a teenager is more likely to engage in a conversation about video games than in one about retirement savings; social events are more likely to be a frequent topic of conversation for extroverted individuals than for introverts. Conversely, while language carries information about who has produced it, knowing who has produced a given utterance can help decode its meaning. An abstract word like "freedom" can mean very different things if uttered by a convict versus a Republican US senator, and ironic statements generally require some background knowledge on the speaker to be understood as such.

The relationship between language and individual differences has been investigated in multiple fields. Previous research has uncovered systematic associations between patterns of language use and demographics (Bamman et al., 2012; Liesenfeld et al., 2021), personality traits (Christian et al., 2021; Ireland and Mehl, 2014; Park et al., 2015; Schwartz et al., 2013; Yarkoni, 2010), mood disorders (Eichstaedt et al., 2018; Tackman et al., 2019; Schwartz et al., 2014), conditions such as schizophrenia (Elvevåg et al., 2011; de Boer et al., 2020; Li et al., 2021; Mitchell et al., 2015; Parola et al., 2022) or ASD (Boorse et al., 2019; Rouhizadeh et al., 2014; Song et al., 2021), and political affiliation (Tatman et al., 2017). Conversely, conditioning text representations on author attributes (e.g., gender or personality) has been shown to enhance performance in several NLP tasks, ranging from sentiment analysis to stance detection (Bamman and Smith, 2015; Flek, 2020; Hovy, 2018, 2015; Lynn et al., 2019). Conditioning on author attributes and states can also improve performance in language modeling and generation (Harrison et al., 2019; Oba et al., 2019; Oraby et al., 2018; Soni et al., 2022; Welch et al., 2020).

However, most research in these domains has focused on developing predictive methods to infer individual attributes from text, or on investigating how conditioning text representations on such attributes can improve performance in downstream NLP tasks. Little work (Wu et al., 2020) has been devoted to exploring self-supervised approaches to language-based author encoding, where compre-

Findings of the Association for Computational Linguistics: EMNLP 2022, pages 1701–1714
December 7-11, 2022 ©2022 Association for Computational Linguistics

hensive representations of authors' profiles - with applications in both author attribute prediction and contextualization - are inferred from unlabelled text. Self-supervised approaches to author encoding would yield two significant advantages over supervised methods. First, these do not rely on the availability of labelled data. Secondly, models trained through self-supervised methods may learn to describe users along dimensions of individual variation which are not captured by standard descriptors such as demographics and personality.

In this paper, we present a self-supervised approach to fine-tuning transformers as language-based user[1] encoders. Building on insights from transfer learning (Ruder et al., 2019) and contrastive representation learning (Gao et al., 2021; Rethmeier and Augenstein, 2021; Xie et al., 2022), we fine-tune a pretrained DistilBERT architecture (Sanh et al., 2020) on a variant of triplet loss (Schroff et al., 2015) using Reddit submissions from more than 1.7m users. Our training objective incentivizes the model to maximize similarity between aggregate representations of posts produced by the *same* author, and to minimize similarity between representations of posts produced by *different* authors. This objective tunes models to detecting linguistic signatures of individuals. If, as suggested by previous studies, linguistic styles carry information about individual traits, this will result in models implicitly learning to extract rich information about user characteristics. Models trained through this contrastive approach could be deployed for text-based prediction of individuals attributes and related behaviors - with potentially impactful applications in language-based clinical and psychological assessment (see Chekroud et al. 2021; Zhang et al. 2022) - or for author-based conditioning in the context of contextualized text classification, language modeling and language generation (Flek, 2020; Hovy, 2018; Kanwal et al., 2021; Leung et al., 2020; Ma et al., 2011; Oba et al., 2019).

In this paper, we describe the training methodology and analyze the behavior of trained models through a battery of intrinsic analyses and probing tasks. These analyses are designed to shed light on: a) which linguistic signatures models rely on in the contrastive task; b) which author attributes are encoded in their representations, and how they vary as a function of training parameters (number of input posts); c) whether learned user representations yield performance advantages in downstream tasks relative to standard pretrained models.

To encourage experimentation and evaluation on additional tasks and datasets, we make our models available on the Hugging Face model hub: see https://huggingface.co/rbroc/contrastive-user-encoder-multipost (multi-anchor) and https://huggingface.co/rbroc/contrastive-user-encoder-singlepost (single-anchor). We also share our code on GitHub: https://github.com/rbroc/contrastive-user-encoders.

2 Contrastive learning

2.1 Task details

We fine-tune a pretrained DistilBERT model on triplet loss, a contrastive learning function first introduced in the context of face encoding (Schroff et al., 2015). In triplet loss, models are fed a triplet of inputs: an anchor a – i.e., an image depicting the face of a given individual; a positive example p – i.e., a different image depicting the same face; and a negative example n – i.e., an image depicting a different face. The three inputs are encoded into high-dimensional embeddings $f(a)$, $f(p)$, $f(n)$, which are used to compute the loss:

$$max(||f(a) - f(n)|| - ||f(a) - f(p)|| + \alpha, 0)$$

where α is a tunable parameter called margin. This function incentivizes models to produce similar embeddings for images of the same face and different embeddings for images of different faces. To do so, the model must learn to detect features of faces that are generally helpful to describe and identify faces and carry this information over to its output encodings.

We transfer this approach to text to train a language-based author encoder – that is, a model that learns to produce compact representation of an individual based on her linguistic behavior, with downstream applications in language-based prediction of individual attributes and contextualized and personalized NLP. To do so, we train DistilBERT on triplets consisting of: a) a set of Reddit submissions from a given user (the anchor, A); b) another Reddit post from the same user (the positive example p); c) a Reddit post from a different, randomly selected user (the negative example n).

To compute the loss, we use [CLS] encodings of the anchors, positive examples and negative exam-

[1]We use *user* and *author* interchangeably.

ples from the last layer of the DistilBERT encoder. We experiment with two training protocols: a) fine-tuning on triplets containing one anchor only (as in traditional triplet loss training); b) fine-tuning on triplets containing up to 10 anchor posts (depending on the number of total posts available for each user) using the feature-wise average of anchor encodings (see Figure 4 in Supplementary Materials for a visual illustration) to compute the loss.

To facilitate interpretation, we evaluate the performance of trained models and baselines on the following accuracy metric (henceforth: contrastive attribution accuracy), which quantifies how often the distance between posts from the same user is lower than the distance between encodings of different users. For a given triplet $t = \{A, p, n\}$, accuracy a_t is calculated as:

$$a_t = \begin{cases} 1, & \text{if } ||\overline{f(A)} - f(n)|| > ||\overline{f(A)} - f(p)|| \\ 0, & \text{if } ||\overline{f(A)} - f(n)|| \leq ||\overline{f(A)} - f(p)|| \end{cases}$$

Conceptually, this metric expresses the model's ability to correctly identify which of two randomly sampled posts p and n belongs to the same author as A based uniquely on their relative proximity to A in embedding space.

We expect that training on a single anchor *versus* on aggregate representations of multiple anchors would not only yield higher contrastive accuracy (as more text is provided), but also allow models to focus on more stable and robust linguistic markers, and facilitate abstraction of higher-level psychological and personality attributes.

2.2 Database

Datasets for both the contrastive learning task and for downstream tasks are constructed from a large-scale database which includes all Reddit submissions in English produced between 2018 and 2019, and authored by users who have posted at least 5 times in that time frame and in at least 5 different subreddits. This amounts to 35m submissions and to more than 1.7m unique users. We created this database by downloading all relevant submissions from Pushshift (`https://pushshift.io`), and filtering along the above-mentioned criteria.

2.3 Triplet dataset

We generate the dataset for triplet loss training as follows. First, we randomly sample one post per user from the database. This set of left-out posts (N) will be used to sample negative examples. Secondly, for each user u in the database, we construct a triplet $T_u = \{A, p, n\}$ by:

- Randomly sampling from the database one post authored by u and using this as the positive example p;
- Sampling a subset of the remaining posts authored by u, to be used as anchors A. We sample 1 post for single-anchor training, and up to 10 for multi-anchor training;
- Randomly sampling a post from a different user from N and using this as the negative example n.

This results in more than 1.7m triplets (one per unique user), which are split into a training set with around 1.24m triplets, a validation set with around 300k triplets, and a test set with around 167k triplets. All posts are tokenized using the pretrained DistilBERT tokenizer (*distilbert-base-uncased* on *transformers*). We use max-length truncation (512 tokens) and padding to generate examples of equal length. Each post in the database is only used once.

2.4 Models

We initialize the DistilBERT model from the English pretrained model *distilbert-base-uncased* available on *transformers* (Wolf et al. 2020, see model card at `https://huggingface.co/distilbert-base-uncased`). We wrap its Tensorflow implementation into a custom model class that allows simultaneous encoding and subsequent aggregation of multiple posts. Our implementation supports model initialization from all model classes and checkpoints available on *transformers*, encouraging reuse and experimentation. The code also makes it easy to add linear or variational compression heads on top of the encoder and experiment with output encodings of varying dimensionality. We compare performance of the fine-tuned models to pretrained DistilBERT and to bag-of-words and Word2Vec (SpaCy implementation, *en_core_web_md*) baselines. We instantiate multiple bag-of-words models varying in the type of representation used (frequencies, word counts, or binary indicators), in dimensionality (100, 1000 or 5000 tokens) and in the distance metric used to compute contrastive attribution accuracy (Euclidean vs. Manhattan distance).

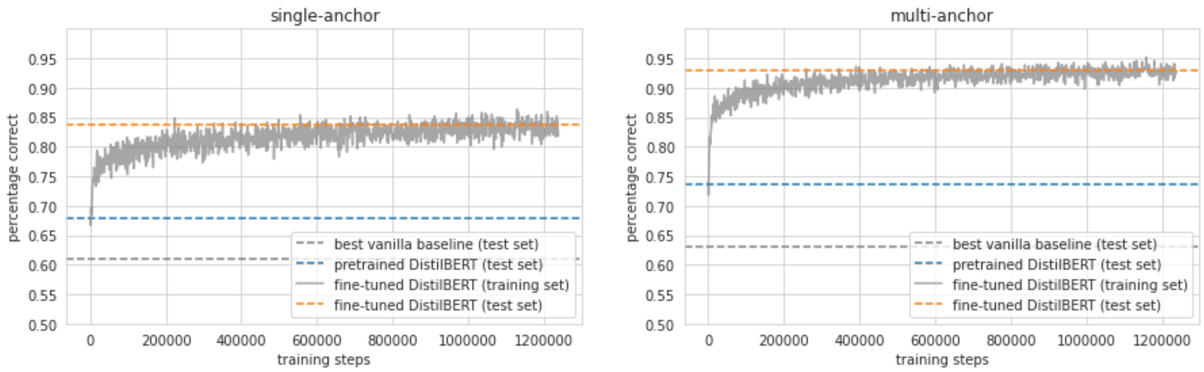

Figure 1: Model performance on training and test set during the first epoch of training in both the single-anchor and the multi-anchor setup. Performance is computed as the proportion of triplets in the dataset for which the distance between encodings of the anchor(s) and of the positive example is lower than the distance between encodings of the anchor(s) and of the negative example (*contrastive attribution accuracy*).

2.5 Training details

For optimization, we use Adam with weight decay (Kingma and Ba, 2017), initialized using the following parameters: initial learning rate: 2e-5; 10k warm-up steps; weight decay rate: 0.01; β_1=.9, β_2=.999, ϵ=1e-6. We use the Keras implementation available at `https://github.com/google-research/bert/blob/master/optimization.py`. We tune the triplet loss margin α and set it to 1. Due to the high memory requirements of simultaneously encoding multiple posts and to the constraints of our training infrastructure (4 GeForce RTX3070 GPUs, 8Gb each), we train the model on small (4-triplet) mini-batches. We do not use gradient accumulation as initial experimentation did not yield significant differences in performance compared to single-batch updates.

3 Evaluation

In this section, we evaluate models and probe their heuristics through a series of intrinsic analyses. First, we simply compare the contrastive attribution accuracy of fine-tuned models with that of baseline models and of pretrained DistilBERT, to evaluate to what extent self-supervised fine-tuning improves models' ability to identify markers of individual styles across the two training scenarios. Secondly, we evaluate model performance as a function of the length of anchor posts, to investigate whether models' heuristics rely on sentence-level stylistic markers or on abstract signatures of individuals' profiles. If models rely on detecting specific sentence-level markers (e.g., particular lexical choices or syntactic constructions), performance should decrease as the length of input posts decreases. On the other hand, if models rely on robust, abstract signatures of user characteristics, performance should be significantly less affected by length. Thirdly, we investigate the role of semantics in models' heuristics. If models rely overwhelmingly on semantics to encode and identify users, performance should be low for triplets where there is little semantic overlap between the anchor(s) and the positive example. Finally, we analyze how attention weights for a large number of vocabulary tokens change between pre-trained and fine-tuned models, in order to gain qualitative insights on whether and how implicit encoding of user characteristics is enhanced by contrastive fine-tuning.

3.1 Implicit evaluation

The best vanilla baseline (Word2Vec with Euclidean distance) achieves .61 accuracy in the single-anchor training and .63 in the multi-anchor training. Pretrained DistilBERT outperforms all vanilla baselines in both scenarios, achieving .68 and .74 accuracy respectively. Fine-tuning for one epoch significantly improves the contrastive attribution performance. Performance increases to .84 in the single-anchor scenario, and it reaches .93 in the multi-anchor scenario (Figure 1). Training for additional epochs does not improve validation performance.

3.2 Sentence-level vs. complex markers

To better understand whether model heuristics rely on detecting individual sentence-level stylistic markers (e.g., lexical patterns or syntactic constructions) or more complex signatures of individuals'

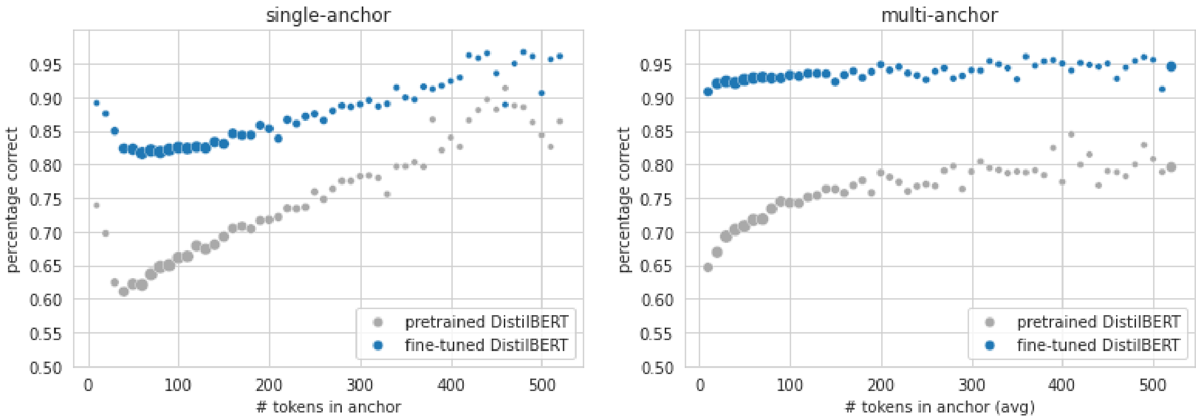

Figure 2: Model performance on the test set as a function of the average number of tokens in anchor posts. Examples are binned into 5-token bins. The size of the dots represents the number of examples per bin.

profiles, we analyze performance as a function of length of anchor posts. The rationale for this is that models relying on multiple and more abstract linguistic signatures should perform well regardless of the length of the input, while accuracy should decrease significantly as sequence length decreases if models tend to rely on detecting specific low-level markers of linguistic styles.

Figure 2 displays model performance as a function of the average length of anchor posts (number of tokens), for both the single-anchor and the multi-anchor model. Performance varies substantially for the single anchor model (range: .81 - .96), while the multi-anchor model is only moderately affected by variation in input length (range: .91 - .96). Interestingly, this is different from what we observe for pretrained DistilBERT, where performance varies substantially as a function of input length both in the single-anchor (range: .61 - .92) and in the multi-anchor scenario (range: .64 - .85). This suggests that training on aggregate representations of multiple posts reduces dependency on local characteristics of individual input sequences, and allows the model to learn to efficiently extract more robust abstract representations of users' language.

3.3 Semantic overlap

To recognize whether two posts have been authored by the same user, models may simply learn to rely on overlaps in semantic information between them. To evaluate the contribution of semantics to models' heuristics, we visualize performance as a function of whether the positive example comes from the same subreddit of at least one of the anchors. If models' heuristics are overwhelmingly based on semantics, accuracy should be low for triplets where there is no overlap.

We observe that (see Table 1), both for the single-anchor model and for the multi-anchor model, accuracy is higher when at least one anchor comes from the same subreddit as the positive example, but performance for no-overlap triplets is far from catastrophic. For the single-anchor model, accuracy is .83 for no-overlap triplets *versus* .92 for triplets with overlap, while for multi-anchor training, accuracy is .91 for no-overlap triplets *versus* .95 for triplets with overlap. These values are also considerably higher than accuracies for pretrained DistilBERT, and performance differences across the two types of triplets are much larger for the pretrained model. This suggests that contrastive learning allows models to develop heuristics that are less dependent on semantics and arguably a complex combination of multiple facets of individual styles.

3.4 Qualitative trait analysis through token-wise attention differentials

To gain further insights on whether learning to encode complex markers of individual styles implicitly leads to encoding information about authors' profiles, we analyze how attention weights for a wide range of vocabulary tokens change between pre-trained and fine-tuned models. Tokens which are consistently assigned higher attention weight after fine-tuning – and which therefore contribute more to aggregate post- and user-level representations after fine-tuning - may reveal qualitative information on which individual traits models learn to infer based on users' text.

To investigate this, we extract context-independent attention differentials for a large set

# anchors	Pretrained no overlap	Pretrained, overlap	Fine-tuned, no overlap	Fine-tuned, overlap
1	.66	.80	.83	.92
10	.71	.76	.91	.95

Table 1: Model performance as a function of overlap between the subreddit of the positive example and the subreddit(s) of the anchor(s).

of vocabulary tokens. We sample 10k posts from the training set, and for each vocabulary token t and each model m (pretrained DistilBERT and the two fine-tuned models), we extract a matrix $A(m, t)$ which contains all attention weights with the [CLS] token as the key and the positional index at which t occurs as the query. We then average all values in the matrix to extract an aggregate context-independent attention score $a_{m,t}$ which quantifies the overall attention score for the token t[2]. Attention differentials for each token are computed by subtracting its attention score in the target fine-tuned model with its attention score in the pretrained model. These values describes how the influence of each token on aggregate representations changes with fine-tuning, with positive values indicating a stronger influence on model representations and negative values indicating a weaker influence.

Table 2 displays the 50 tokens with largest positive attention differentials for both the single-anchor and multi-anchor models. For the single-anchor model, many tokens with large positive differentials are strongly marked for gender, age, ethnicity, or political views (e.g., husband, bride, breast, boyfriend, uncle, nephew, japanese, euro, abortion, army). No clear pattern emerges among tokens with negative differentials (see Table 3 in Supplementary Materials). Multi-anchor models feature many potential mental health or personality indicators among tokens with top attention differentials (e.g., suicidal, gambling, desperately, worthless, obsessed, abusive), suggesting that training on aggregate representations of multiple tokens may facilitate abstraction of higher-order traits (in line with our previous analyses). While these are obviously qualitative interpretations that call for further systematic investigations (e.g., predictive validation on labeled datasets), they corroborate the hypothesis that, after fine-tuning, models place higher emphasis on linguistic patterns that are indicative of individual traits along multiple axes of variation.

4 Classification tasks

We also evaluate models on a battery of probing tasks designed to test whether fine-tuning yields performance advantages on user-based predictive tasks. Probing tasks are designed as follows. For each of the 30 most popular subreddits in the database, we train a classifier to predict whether a given user has posted at least once in that subreddit based uniquely on posts produced by the same user in *unrelated* subreddits. Since no reference post from the target subreddit is provided and choice of input posts is in principle asystematic, performing this task relies on models being able to infer useful and generalizable information on users' profiles from their posts, and to use it to predict unknown preferences and behaviors. If fine-tuning facilitates this process, performance of fine-tuned models should be consistently higher than performance of pretrained models.

For each subreddit, we build a training, a validation and a test dataset. To build the dataset for a given subreddit s, we first extract from the database the ids of all users who have posted in s at least once. For each user, we sample up to 10 submissions drawn from other subreddits. We then sample an equal number of users who have never posted in s, and extract up to 10 posts for each of them. We split the resulting dataset into 75/15/15 training/validation/test splits.

Aggregate encodings[3] of users' posts are used as inputs to simple classifiers (one per model and subreddit), which are optimized to predict whether a given user has posted in s at least once based uniquely on these. All classifiers are trained for 3 epochs. Similar to the contrastive learning task, we

[2]To provide a concrete example, the context-independent attention score for the token "house" is computed by averaging attention weights for all ([CLS], "house") key/value pairs in the 10000-posts corpus.

[3]For DistilBERT models, aggregation is performed by computing the feature-wise average of [CLS] encodings, as for anchors in the contrastive training task. For vanilla baselines, aggregation is performed by averaging the bag-of-words or Word2Vec representations across all posts.

10-anchor model	1-anchor model
offers, obsessed, crypt, xiao, sponge, leaked, suicidal, keen, pathetic, https, diverse, downloaded, ##lika, psychedelic, tran, purchasing, gambling, tents, banned, desperately, breed, bribe, ##grapher, ugly, wits, ##folk, divorced, tia, ##km, bc, abusive, folks, trance, worthless, wanna, husband, tee, disco, karma, jungle, rory, nyc, toni, probation, buddy, inexpensive, quantity, ##tive, http, prom, brass, bois, mir, fraternity, encouraging, tempered, [SEP], cheers	nephew, perfection, army, dated, abortion, lads, ##rang, uncle, wireless, nyc, historically, dashboard, comrade, article, ##riation, trained, japanese, profession, daddy, journalist, title, scientists, kidding, thanksgiving, albuquerque, hacker, bard, euro, shane, jai, roman, beautifully, lipstick, linear, ##pile, ito, pee, fragrance, width, tia, neighbors, rig, united, unpopular, bride, lease, ##rse, margarita, buffy, husband, toni, linux, ##ame, pardon, aaa, notebook, [SEP], boyfriend, breast, fiance

Table 2: Top-50 tokens with highest and attention differentials for 10-anchor and 1-anchor model.

experiment with two training protocols, differing in the number of input posts (one vs. up to 10). Note that, for DistilBERT models, only the classification head is tuned.

Figure 3 displays classification accuracy for all subreddits. Results corroborate our predictions. Fine-tuned models perform better than all baselines in all classification tasks, both in the single-post and in the multi-post scenario. Performance gains relative to pretrained DistilBERT are larger in the single-post scenario than in the multi-post scenario, arguably reflecting the fact that availability of multiple posts increases chances of topic overlap with the target subreddit, thus increasing the effectiveness of semantics-based heuristics and reducing the need for user-based ones.

Note that the magnitude of performance gains varies widely across subreddits (see figure 5 in the Supplementary Materials for details). Performance differences are large for subreddits where likelihood to participate is arguably influenced by personality or demographics (e.g., *teenagers*, *Jokes*, *relationship_advice*), but we also observe moderate to large gains for subreddits focused on specific video games (e.g., *RocketLeagueExchange*). This may reflect some degree of overfitting to Reddit-specific discourse. Training a truly generalizable language-based author encoder in a self-supervised fashion will require training on data drawn from *multiple* sources, thus avoiding overrepresentation of platform-specific axes of individual variation in e.g., topics, styles and demographics.

5 Discussion

We introduced a self-supervised approach to training generalized language-based author encoders. We showed that models fine-tuned on user-based triplet loss learn to infer generalizable information on user profiles from complex patterns of linguistic behavior.

Author encoders may have impactful applications in a variety of domains. Two particularly important examples are language-based clinical, psychological and personality assessment (Skaik and Inkpen, 2020). Language-based encodings of individuals could potentially be used to predict personality, psychological traits or even clinical diagnoses and symptoms from spontaneous text - especially for disorders, such as depression, that have previously been associated with consistent patterns of linguistic behavior. For both psychological and clinical applications, complementing traditional methods with naturalistic text-based techniques could not only yield general performance advantages, but also help increase scalability and generalizability (Panch et al., 2020; Parola et al., 2022; Rybner et al., 2022), and reduce subjective biases (Park et al., 2015). At the moment, however, no large-scale datasets are publicly available which make it possible to benchmark our models on tasks relevant to these applications. Given the large potential for positive societal impact, we believe that the NLP community should promote interdisciplinary efforts aimed at collecting and safely sharing such resources (Chekroud et al., 2021; Dukart et al., 2021; Ewbank et al., 2020; Low et al., 2020).

Language-based user encodings learned through self-supervised methods could also have a significant impact on contextualized and personalized NLP (Flek, 2020). Previous work has shown that conditioning representations of text sequences on author traits is beneficial for both downstream tasks and language modeling and generation. Contextualization through user embeddings encoding rich

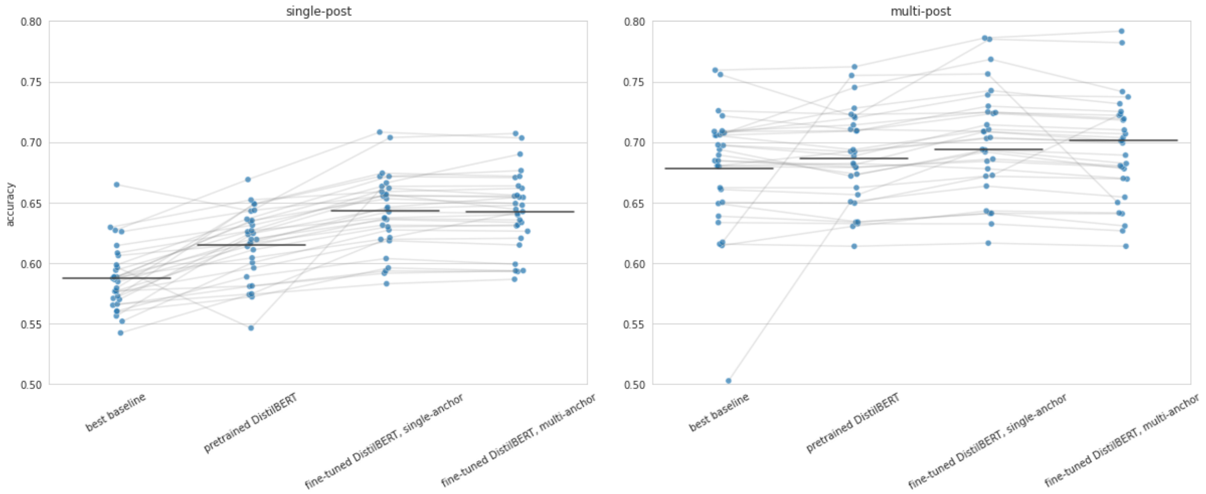

Figure 3: Classification performance for best vanilla baseline, pretrained DistilBERT and the fine-tuned models for each of the 30 target subreddit. Bars represent mean performance across subreddits.

information about user characteristics and their linguistic styles could prove a valid alternative, especially for tasks – such as irony or stance detection – that are particularly challenging in absence of complex contextual information (Hovy and Yang, 2021; Lynn et al., 2017). As a follow-up to this work, we are currently exploring the effect of author-based self-supervised contextualization during pre-training of bidirectional transformers.

6 Conclusion

In this paper, we presented a self-supervised approach to training a language-based user encoder. We showed that models trained on user-based triplet loss can learn compact user representations that encode information about individual traits and yield performance benefits in downstream user-based tasks. Future directions include scaling this approach to larger and more diverse data, developing resources for direct evaluation on societally important predictive tasks (e.g., psychiatric assessment), and exploring their potential to empower novel approaches to contextualized natural language modeling, understanding and generation.

7 Limitations

Our work introduces a self-supervised approach to training a generalized language-based author encoder. We highlighted important applications in clinical and contextualized and personalized NLP. This paper is intended to lay the methodological foundations for these applications, which will be explored directly in future work.

Being exclusively trained on Reddit data, our models probably overfit to linguistic markers and traits which are relevant to characterizing the Reddit user population, but less salient in the general population (e.g., video games preferences). Training on more and more diverse data (i.e., from multiple discourse types and a broader population) will be required to train a truly "universal" user encoder.

Furthermore, our self-supervised approach enforces little or no control over biases, which models may actively use as part of their heuristics in contrastive and downstream tasks (Bender et al., 2021; Davidson et al., 2019; Mitchell et al., 2019; Ferrer et al., 2020; Koolen and van Cranenburgh, 2017; Xia et al., 2020; Zhou et al., 2021; Hovy and Spruit, 2016). Future iterations will require the implementation of thorough bias testing and, potentially, the introduction of optimization constraints at training that help counter their emergence (Shah et al., 2020).

Finally, it is important to highlight that models may be used for malicious applications such as identifying and targeting social media users. Mapping and discussing these risks within the landscape of current data and AI regulatory frameworks is central to future developments of this line of work.

Acknowledgements

We thank Alejandro de la Vega, Riccardo Fusaroli, James D. Kent, Ross Blair, and Marco Carapezza for their helpful comments.

References

David Bamman, Jacob Eisenstein, and Tyler Schnoebelen. 2012. Gender in Twitter: Styles, stances, and social networks.

David Bamman and Noah Smith. 2015. Contextualized Sarcasm Detection on Twitter. *Proceedings of the International AAAI Conference on Web and Social Media*, 9(1):574–577. Number: 1.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pages 610–623, New York, NY, USA. Association for Computing Machinery.

Jaclin Boorse, Meredith Cola, Samantha Plate, Lisa Yankowitz, Juhi Pandey, Robert T. Schultz, and Julia Parish-Morris. 2019. Linguistic markers of autism in girls: evidence of a "blended phenotype" during storytelling. *Molecular Autism*, 10(1):14.

Adam M. Chekroud, Julia Bondar, Jaime Delgadillo, Gavin Doherty, Akash Wasil, Marjolein Fokkema, Zachary Cohen, Danielle Belgrave, Robert DeRubeis, and Raquel Iniesta. 2021. The promise of machine learning in predicting treatment outcomes in psychiatry. *World Psychiatry*, 20(2):154–170. Publisher: Wiley Online Library.

Hans Christian, Derwin Suhartono, Andry Chowanda, and Kamal Z. Zamli. 2021. Text based personality prediction from multiple social media data sources using pre-trained language model and model averaging. *Journal of Big Data*, 8(1):68.

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial Bias in Hate Speech and Abusive Language Detection Datasets. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 25–35, Florence, Italy. Association for Computational Linguistics.

Janna N. de Boer, Sanne G. Brederoo, Alban E. Voppel, and Iris E. C. Sommer. 2020. Anomalies in language as a biomarker for schizophrenia. *Current Opinion in Psychiatry*, 33(3):212–218.

Juergen Dukart, Susanne Weis, Sarah Genon, and Simon B. Eickhoff. 2021. Towards increasing the clinical applicability of machine learning biomarkers in psychiatry. *Nature Human Behaviour*, 5(4):431–432. Number: 4 Publisher: Nature Publishing Group.

Johannes C. Eichstaedt, Robert J. Smith, Raina M. Merchant, Lyle H. Ungar, Patrick Crutchley, Daniel Preoţiuc-Pietro, David A. Asch, and H. Andrew Schwartz. 2018. Facebook language predicts depression in medical records. *Proceedings of the National Academy of Sciences*, 115(44):11203–11208. Publisher: Proceedings of the National Academy of Sciences.

B. Elvevåg, K. Helsen, M. De Hert, K. Sweers, and G. Storms. 2011. Metaphor interpretation and use: A window into semantics in schizophrenia. *Schizophrenia Research*, 133(1):205–211.

Michael P. Ewbank, Ronan Cummins, Valentin Tablan, Sarah Bateup, Ana Catarino, Alan J. Martin, and Andrew D. Blackwell. 2020. Quantifying the Association Between Psychotherapy Content and Clinical Outcomes Using Deep Learning. *JAMA Psychiatry*, 77(1):35–43.

Xavier Ferrer, Tom van Nuenen, Jose M. Such, and Natalia Criado. 2020. Discovering and Categorising Language Biases in Reddit. Technical Report arXiv:2008.02754, arXiv. ArXiv:2008.02754 [cs] type: article.

Lucie Flek. 2020. Returning the N to NLP: Towards Contextually Personalized Classification Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7828–7838, Online. Association for Computational Linguistics.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*.

Vrindavan Harrison, Lena Reed, Shereen Oraby, and Marilyn Walker. 2019. Maximizing Stylistic Control and Semantic Accuracy in NLG: Personality Variation and Discourse Contrast. Technical Report arXiv:1907.09527, arXiv. ArXiv:1907.09527 [cs] type: article.

Dirk Hovy. 2015. Demographic Factors Improve Classification Performance. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 752–762, Beijing, China. Association for Computational Linguistics.

Dirk Hovy. 2018. The Social and the Neural Network: How to Make Natural Language Processing about People again. In *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, pages 42–49, New Orleans, Louisiana, USA. Association for Computational Linguistics.

Dirk Hovy and Shannon L. Spruit. 2016. The Social Impact of Natural Language Processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598, Berlin, Germany. Association for Computational Linguistics.

Dirk Hovy and Diyi Yang. 2021. The Importance of Modeling Social Factors of Language: Theory and Practice. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 588–602, Online. Association for Computational Linguistics.

Molly E. Ireland and Matthias R. Mehl. 2014. Natural Language Use as a Marker of Personality. ISBN: 9780199838639.

Safia Kanwal, Sidra Nawaz, Muhammad Kamran Malik, and Zubair Nawaz. 2021. A Review of Text-Based Recommendation Systems. *IEEE Access*, 9:31638–31661. Conference Name: IEEE Access.

Diederik P. Kingma and Jimmy Ba. 2017. Adam: A Method for Stochastic Optimization. Technical Report arXiv:1412.6980, arXiv. ArXiv:1412.6980 [cs] type: article.

Corina Koolen and Andreas van Cranenburgh. 2017. These are not the Stereotypes You are Looking For: Bias and Fairness in Authorial Gender Attribution. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 12–22, Valencia, Spain. Association for Computational Linguistics.

John Kalung Leung, Igor Griva, and William G. Kennedy. 2020. Text-based Emotion Aware Recommender. In *Computer Science & Information Technology*, pages 101–114. ArXiv:2007.01455 [cs].

Chuyuan Li, Maxime Amblard, Chloé Braud, Caroline Demily, Nicolas Franck, and Michel Musiol. 2021. Investigating non lexical markers of the language of schizophrenia in spontaneous conversations. In *Proceedings of the 2nd Workshop on Computational Approaches to Discourse*, pages 20–28, Punta Cana, Dominican Republic and Online. Association for Computational Linguistics.

Andreas Liesenfeld, Gábor Parti, Yu-Yin Hsu, and Chu-Ren Huang. 2021. Predicting gender and age categories in English conversations using lexical, non-lexical, and turn-taking features. Technical Report arXiv:2102.13355, arXiv. ArXiv:2102.13355 [cs] type: article.

Daniel M. Low, Laurie Rumker, Tanya Talker, John Torous, Guillermo Cecchi, and Satrajit S. Ghosh. 2020. Reddit Mental Health Dataset. Type: dataset.

Veronica Lynn, Salvatore Giorgi, Niranjan Balasubramanian, and H. Andrew Schwartz. 2019. Tweet Classification without the Tweet: An Empirical Examination of User versus Document Attributes. In *Proceedings of the Third Workshop on Natural Language Processing and Computational Social Science*, pages 18–28, Minneapolis, Minnesota. Association for Computational Linguistics.

Veronica Lynn, Youngseo Son, Vivek Kulkarni, Niranjan Balasubramanian, and H. Andrew Schwartz. 2017. Human Centered NLP with User-Factor Adaptation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1146–1155, Copenhagen, Denmark. Association for Computational Linguistics.

Hao Ma, Dengyong Zhou, Chao Liu, Michael R. Lyu, and Irwin King. 2011. Recommender systems with social regularization. In *Proceedings of the fourth ACM international conference on Web search and data mining*, WSDM '11, pages 287–296, New York, NY, USA. Association for Computing Machinery.

Margaret Mitchell, Kristy Hollingshead, and Glen Coppersmith. 2015. Quantifying the Language of Schizophrenia in Social Media. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 11–20, Denver, Colorado. Association for Computational Linguistics.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pages 220–229, New York, NY, USA. Association for Computing Machinery.

Daisuke Oba, Naoki Yoshinaga, Shoetsu Sato, Satoshi Akasaki, and Masashi Toyoda. 2019. Modeling Personal Biases in Language Use by Inducing Personalized Word Embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2102–2108, Minneapolis, Minnesota. Association for Computational Linguistics.

Shereen Oraby, Lena Reed, Shubhangi Tandon, T. S. Sharath, Stephanie Lukin, and Marilyn Walker. 2018. Controlling Personality-Based Stylistic Variation with Neural Natural Language Generators. Technical Report arXiv:1805.08352, arXiv. ArXiv:1805.08352 [cs] type: article.

Trishan Panch, Tom J. Pollard, Heather Mattie, Emily Lindemer, Pearse A. Keane, and Leo Anthony Celi. 2020. "Yes, but will it work for my patients?" Driving clinically relevant research with benchmark datasets. *npj Digital Medicine*, 3(1):1–4. Number: 1 Publisher: Nature Publishing Group.

Gregory Park, H. Andrew Schwartz, Johannes C. Eichstaedt, Margaret L. Kern, Michal Kosinski, David J. Stillwell, Lyle H. Ungar, and Martin E. P. Seligman. 2015. Automatic personality assessment through social media language. *Journal of Personality and Social Psychology*, 108(6):934–952. Place: US Publisher: American Psychological Association.

Alberto Parola, Jessica Mary Lin, Arndis Simonsen, Vibeke Bliksted, Yuan Zhou, Huiling Wang, Lana Inoue, Katja Koelkebeck, and Riccardo Fusaroli. 2022. Speech disturbances in schizophrenia: assessing cross-linguistic generalizability of NLP automated measures of coherence. Technical report, medRxiv. Type: article.

Nils Rethmeier and Isabelle Augenstein. 2021. A primer on contrastive pretraining in language processing:

Methods, lessons learned & perspectives. *ACM Computing Surveys (CSUR)*.

Masoud Rouhizadeh, Emily Prud'hommeaux, Jan van Santen, and Richard Sproat. 2014. Detecting linguistic idiosyncratic interests in autism using distributional semantic models. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 46–50, Baltimore, Maryland, USA. Association for Computational Linguistics.

Sebastian Ruder, Matthew E. Peters, Swabha Swayamdipta, and Thomas Wolf. 2019. Transfer Learning in Natural Language Processing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials*, pages 15–18, Minneapolis, Minnesota. Association for Computational Linguistics.

Astrid Rybner, Emil Trenckner Jessen, Marie Damsgaard Mortensen, Stine Nyhus Larsen, Ruth Grossman, Niels Bilenberg, Cathriona Cantio, Jens Richardt Møllegaard Jepsen, Ethan Weed, Arndis Simonsen, and Riccardo Fusaroli. 2022. Vocal markers of autism: assessing the generalizability of machine learning models. Technical report, bioRxiv. Section: New Results Type: article.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv:1910.01108 [cs]*. ArXiv: 1910.01108.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. FaceNet: A Unified Embedding for Face Recognition and Clustering. pages 815–823.

H. Andrew Schwartz, Johannes Eichstaedt, Margaret L. Kern, Gregory Park, Maarten Sap, David Stillwell, Michal Kosinski, and Lyle Ungar. 2014. Towards Assessing Changes in Degree of Depression through Facebook. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 118–125, Baltimore, Maryland, USA. Association for Computational Linguistics.

H. Andrew Schwartz, Johannes C. Eichstaedt, Margaret L. Kern, Lukasz Dziurzynski, Stephanie M. Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin E. P. Seligman, and Lyle H. Ungar. 2013. Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach. *PLOS ONE*, 8(9):e73791. Publisher: Public Library of Science.

Deven Shah, H. Andrew Schwartz, and Dirk Hovy. 2020. Predictive Biases in Natural Language Processing Models: A Conceptual Framework and Overview. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5248–5264. ArXiv:1912.11078 [cs].

Ruba Skaik and Diana Inkpen. 2020. Using Social Media for Mental Health Surveillance: A Review. *ACM Computing Surveys*, 53(6):129:1–129:31.

Amber Song, Meredith Cola, Samantha Plate, Victoria Petrulla, Lisa Yankowitz, Juhi Pandey, Robert T. Schultz, and Julia Parish-Morris. 2021. Natural language markers of social phenotype in girls with autism. *Journal of Child Psychology and Psychiatry*, 62(8):949–960. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/jcpp.13348.

Nikita Soni, Matthew Matero, Niranjan Balasubramanian, and H. Andrew Schwartz. 2022. Human Language Modeling. Technical Report arXiv:2205.05128, arXiv. ArXiv:2205.05128 [cs] type: article.

Allison M. Tackman, David A. Sbarra, Angela L. Carey, M. Brent Donnellan, Andrea B. Horn, Nicholas S. Holtzman, To'Meisha S. Edwards, James W. Pennebaker, and Matthias R. Mehl. 2019. Depression, negative emotionality, and self-referential language: A multi-lab, multi-measure, and multi-language-task research synthesis. *Journal of Personality and Social Psychology*, 116(5):817–834. Place: US Publisher: American Psychological Association.

Rachael Tatman, Leo Stewart, Amandalynne Paullada, and Emma Spiro. 2017. Non-lexical Features Encode Political Affiliation on Twitter. In *Proceedings of the Second Workshop on NLP and Computational Social Science*, pages 63–67, Vancouver, Canada. Association for Computational Linguistics.

Charles Welch, Jonathan K. Kummerfeld, Verónica Pérez-Rosas, and Rada Mihalcea. 2020. Exploring the Value of Personalized Word Embeddings. Technical Report arXiv:2011.06057, arXiv. ArXiv:2011.06057 [cs] type: article.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. HuggingFace's Transformers: State-of-the-art Natural Language Processing. Technical Report arXiv:1910.03771, arXiv. ArXiv:1910.03771 [cs] type: article.

Xiaodong Wu, Weizhe Lin, Zhilin Wang, and Elena Rastorgueva. 2020. Author2Vec: A Framework for Generating User Embedding. Technical Report arXiv:2003.11627, arXiv. ArXiv:2003.11627 [cs, stat] type: article.

Mengzhou Xia, Anjalie Field, and Yulia Tsvetkov. 2020. Demoting Racial Bias in Hate Speech Detection. In *Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media*, pages 7–14, Online. Association for Computational Linguistics.

Xu Xie, Fei Sun, Zhaoyang Liu, Shiwen Wu, Jinyang Gao, Jiandong Zhang, Bolin Ding, and Bin Cui. 2022. Contrastive learning for sequential recommendation. In *2022 IEEE 38th International Conference on Data Engineering (ICDE)*, pages 1259–1273. IEEE.

Tal Yarkoni. 2010. Personality in 100,000 Words: A large-scale analysis of personality and word use among bloggers. *Journal of Research in Personality*, 44(3):363–373.

Tianlin Zhang, Annika M. Schoene, Shaoxiong Ji, and Sophia Ananiadou. 2022. Natural language processing applied to mental illness detection: a narrative review. *npj Digital Medicine*, 5(1):1–13. Number: 1 Publisher: Nature Publishing Group.

Xuhui Zhou, Maarten Sap, Swabha Swayamdipta, Noah A. Smith, and Yejin Choi. 2021. Challenges in Automated Debiasing for Toxic Language Detection. Technical Report arXiv:2102.00086, arXiv. ArXiv:2102.00086 [cs] type: article.

A Supplementary Materials

A.1 Schematic illustration of contrastive learning through triplet loss

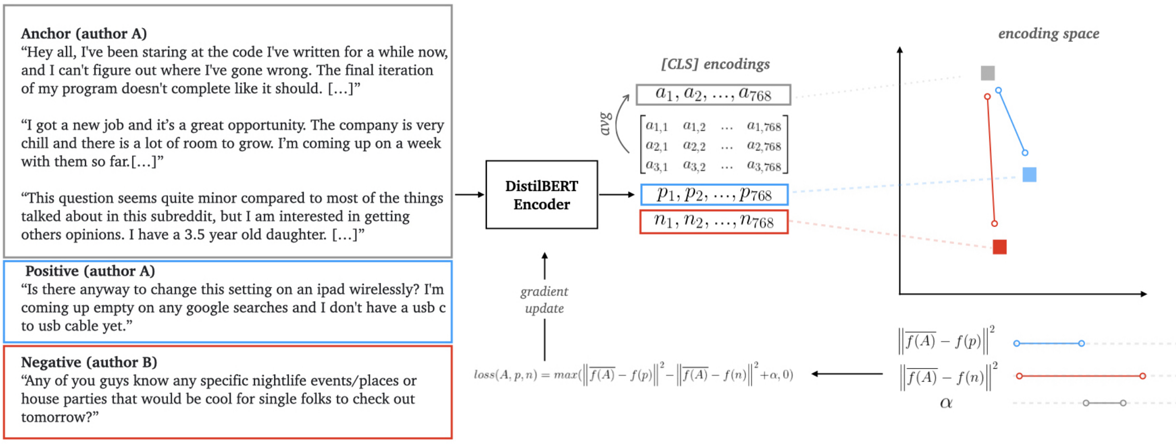

Figure 4: Illustration of contrastive learning through triplet loss. $\overline{f(A)}$ is the feature-wise average of [CLS] encodings for all anchor posts from the last hidden layer of the DistilBERT encode

A.2 Classification performance per subreddit

This is a more detailed breakdown of classification performance for pretrained and fine-tuned models in the downstream classification task (Section 4) for the one-anchor scenario (i.e., when only a single post is used to classify whether a given user has posted in the target subreddit.)

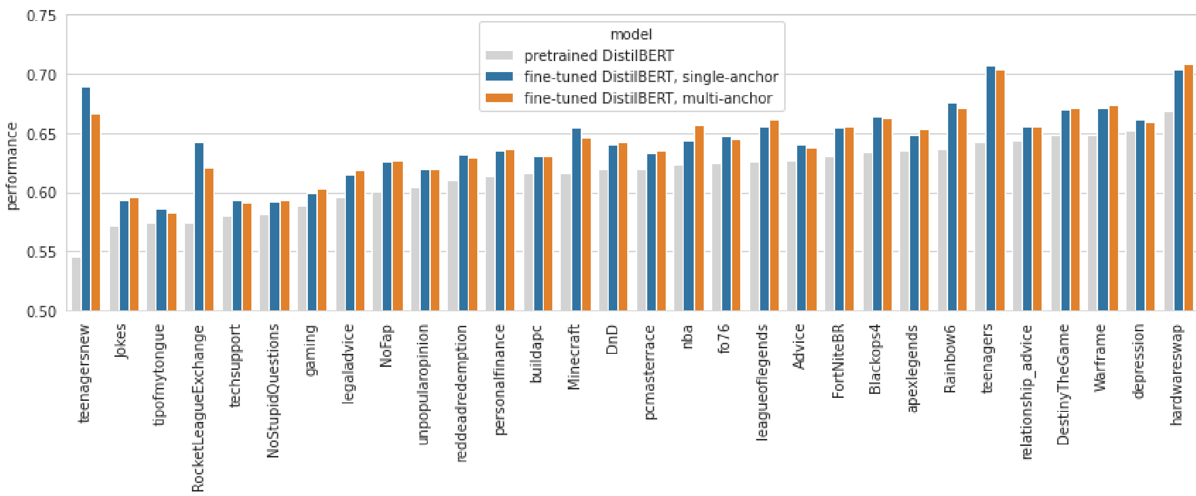

Figure 5: Model performance in downstream tasks per subreddit. Results for classifiers trained on a single input post.

A.3 Attention differentials

The following table shows both the 50 tokens with highest attention differentials, and the 50 tokens with lowest (negative) attention differentials. We discussed patterns in positive attention differentials in the manuscript. Here we highlight that no salient pattern seems to emerge among low-differential tokens.

model	**highest (positive) differentials**	**lowest (negative) differential**
10-anchor model	offers, obsessed, crypt, xiao, sponge, leaked, suicidal, keen, pathetic, https, diverse, downloaded, ##lika, psychedelic, tran, purchasing, gambling, tents, banned, desperately, breed, bribe, ##grapher, ugly, wits, ##folk, divorced, tia, ##km, bc, abusive, folks, trance, worthless, wanna, husband, tee, disco, karma, jungle, rory, nyc, toni, probation, buddy, inexpensive, quantity, ##tive, http, prom, brass, bois, mir, fraternity, encouraging, tempered, [SEP], cheers	terrestrial, skeletons, crossing, cosmetics, milestone, woo, beaver, dam, ghosts, 747, trends, resource, ##ump, ##bber, bs, rune, knuckles, towed, sand, watched, omega, arch, conduct, subjective, ##20, jeopardy, ##ctus, ##gold, bb, induction, ##nton, lit, tricks, knights, ##aca, summon, activate, woods, observation, solos, maia, witch, cookies, rituals, bathroom, tournament, label, odor, spaces, brands, meat, tattoo, depot, titanium, claw, tie, banner, restore, symbol, bounce
1-anchor model	nephew, perfection, army, dated, abortion, lads, ##rang, uncle, wireless, nyc, historically, dashboard, comrade, article, ##riation, trained, japanese, profession, daddy, journalist, title, scientists, kidding, thanksgiving, albuquerque, hacker, bard, euro, shane, jai, roman, beautifully, lipstick, linear, ##pile, ito, pee, fragrance, width, tia, neighbors, rig, united, unpopular, bride, lease, ##rse, margarita, buffy, husband, toni, linux, ##ame, pardon, aaa, notebook, [SEP], boyfriend, breast, fiance	toxin, wheels, boiled, terrestrial, confuse, buzz, ##meo, chickens, repeating, nasty, allergic, streamed, ##ban, karma, knuckles, ##bber, converted, rings, ##oj, consoles, flickering, boil, talent, sequence, scratch, expansion, jerking, submission, quiz, ending, deposit, rumor, corporation, hen, ##ice, replied, locks, aids, donor, shock, remastered, strand, defeated, strengths, chilling, guides, heal, bracket, possibility, grip, ##lot, grim, hatch, superb, adam, healing, collecting, captive, ##gai, brave

Table 3: Tokens with highest and lowest attention differential.