Uso de Transformers para Detecção de Primeiras Impressões em Vídeos

Wallace G. S. Lima, Murillo R. Batista, Josué J. G. Ramos

wallacegsantoslima@gmail.com, {mbatista,jgramos}@cti.gov.br

Divisão de Sistemas Ciberfísicos - DISCF
CTI/MCTI Renato Archer – Campinas/SP

Abstract. *The objective of the project is related to the use of multimodal information to detect personality traits of people through machine learning techniques. The Big Five model was used for personality description; this model is used to help predict behavior as well as personality. From the exploration of datasets, the Chalearn lap 2016/2017 dataset was selected. In conclusion, taking into account the sequential characteristic of the input information, a machine learning model based on the Transformers architecture is proposed for the classification into extroversion and non-extroversion.*

Resumo. *O objetivo do projeto está relacionado ao uso de informações multimodais para detectar traços de personalidade de pessoas através de técnicas de aprendizado de máquina. Utilizou-se o modelo Big Five para a caracterização de personalidade; este é usado para ajudar a prever o comportamento, bem como a personalidade. A partir de um levantamento de datasets, foi selecionado o conjunto de dados Chalearn lap 2016/2017. Por fim, levando em conta a característica sequencial das informações de entrada, propõe-se um modelo de aprendizado de máquinas baseado na arquitetura Transformers para a classificação em extroversão e não-extroversão.*

1. Introdução

O projeto executado tem como base o trabalho *First Impressions: A Survey on Vision-Based Apparent Personality Trait Analysis*[1], no qual os autores discutem os principais aspectos relacionados à predição de traços de personalidade.

Traços de personalidade são um modelo eficiente e muito adotado para a descrição da personalidade de indivíduos; consistem em uma definição e em uma medida, sendo modelos adequados para aplicação de técnicas de aprendizado de máquina e de aprendizado profundo.[1] O modelo de traços adotado no presente estudo foi o *Big Five*, um dos mais utilizados na Psicologia. Ele descreve a personalidade humana em 5 diferentes traços: extroversão, a agradabilidade, a abertura, a conscienciosidade e o neuroticismo[8].

Descrever a personalidade de um indivíduo, a partir deste ângulo, traduz-se em medir o quanto cada traço citado acima o representa. É uma tarefa muito estudada por psicólogos e, até então, a aplicação de questionários é o método mais utilizado para este fim[9]. Devido a isso, a predição da personalidade real, nome dado àquela definida por uma equipe de profissionais de psicologia capacitada e autorizada para tal, torna-se uma tarefa muito difícil. Por conseguinte, muitos trabalhos de estimação de personalidade (*personality computing* em tradução livre) focam na predição da personalidade aparente, ou primeiras impressões: a personalidade de um observado apreendida por um

observador. A personalidade aparente apresenta duas principais características que devem ser mantidas em mente: primeiro, é subjetiva (pertence ao observador) e, segundo, surge durante a primeira interação observador-observado (estudos apontam que pode ser definida em um vislumbre tão curto quanto 100ms).

A predição de primeiras impressões é uma tarefa bastante difícil devido ao caráter subjetivo do julgamento dos dados. Assim, um dos principais desafios da rotulagem dos dados é atribuir uma pontuação para cada traço a partir das análises feitas pelos observadores. Não há padrão para a anotação dos dados. Dependendo da técnica de rotulagem, os resultados podem mudar bastante. A comparação emparelhada (comparar cada imagem em algumas dimensões ao invés de classificá-las individualmente) parece ser um bom método para a redução de tendências/preconceitos na rotulagem de dados. A escolha de um conjunto de dados para o treinamento de um modelo é o desafio inicial.

2. Escolha de Dataset

Considerando os aspectos supracitados, estudou-se os conjuntos de dados apresentados na *survey*[1] ,e, para a realização de experimentos, o conjunto de dados selecionado para o trabalho foi o *Chalearn lap 2016/2017*. Este conjunto consiste em um total de 10.000 clipes curtos, de cerca de 15 segundos cada, coletados a partir de vídeos no *YouTube*. Para cada clipe, foram atribuídos rótulos contínuos no intervalo de 0 a 1 para cada um dos traços de personalidade do *Big Five* e uma pontuação *interview*, de 0 a 1, que indica a probabilidade do indivíduo do vídeo ser convocado para uma entrevista de emprego. Além dos clipes, há também anotações de gênero (*Male* ou *Female*) e de etnia (*Asian*, *Caucasian* ou *African-American*) dos indivíduos presentes nos vídeos. Estas últimas anotações foram utilizadas para realizar investigações de vieses no *dataset*[6].

Os clipes do *dataset* foram coletados a partir de quase três mil canais do *YouTube* diferentes. De cada vídeo, foram extraídos clipes curtos, com um máximo de seis clipes por vídeo, que foram utilizados no *dataset*.

A seguir, cada clipe passou por um processo de rotulagem. É importante ressaltar que os vídeos que foram utilizados como fonte dos clipes do dataset passaram por um processo de filtragem. Em resumo, só foram usados vídeos que cumpriam os seguintes requisitos:

- Uma única pessoa deve aparecer no vídeo;
- Boa qualidade de áudio e imagem;
- Somente inglês;
- Somente pessoas acima de 13/15 anos;
- Pouca movimentação de câmera;
- Sem conteúdo adulto, violento ou nudez;
- Pode ter pessoas ao fundo além do falante, mas devem ser facilmente discerníveis do mesmo;
- Sem propagandas;
- Sem mudanças abruptas de imagem ou de áudio.

Após essa etapa de seleção dos vídeos e da extração, os clipes passam por um processo de rotulagem baseado em *crowdsourcing*. Para isso, foi utilizada a plataforma online *Amazon Mechanical Turk* (AMT), com múltiplos votos por vídeo. Os votos foram feitos por meio de comparação emparelhada e a pontuação final de cada rótulo é definida por um modelo BTL[3]. Um aspecto importante para o processo de rotulagem é que os pares de vídeos são selecionados por um algoritmo de *Small-World*[4], que geram grafos com alta conectividade, evitando regiões que não estejam conectadas.

Cada participante recebe um par de vídeos e, após assisti-los, deve votar em cada um dos aspectos que será rotulado (Figura 1).

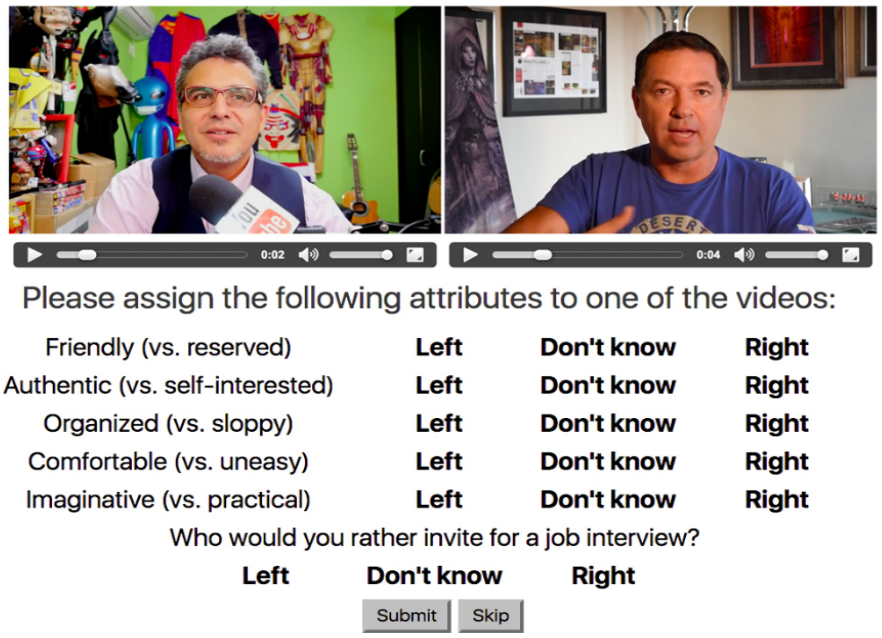

Figura 1. Interface de avaliação dos vídeos.[2]

Por fim, com os resultados da votação, a cada um dos traços é atribuída uma pontuação de 0 a 1, como nos exemplos da Figura 2.

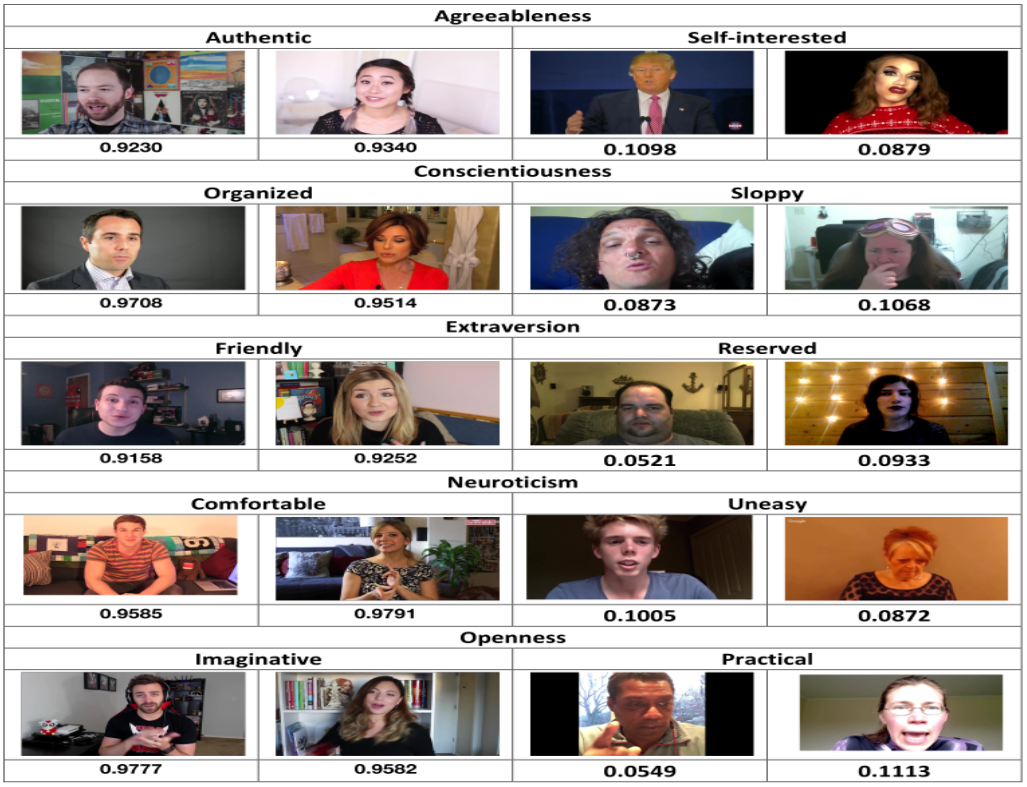

Figura 2. Exemplos de pontuações altas e baixas para cada um dos traços de personalidade do *Big Five*.[2]

Apesar de críticas levantada pela coorientadora e pela orientanda de psicologia do projeto sobre o *dataset* (principalmente no que se refere à pergunta apresentada aos votantes: comentou-se que a descrição apresentada dos traços não necessariamente condiz com os traços em si), optou-se por trabalhar com estes dados, pois é o maior conjunto de dados com entradas multimodais sobre o tema até o presente momento. Assim, a ideia é estudar modelos de aprendizado de máquina que, a partir de vídeos, retornem uma descrição da personalidade aparente de um indivíduo seguindo o modelo *Big Five*.

3. Proposta e Implementação de um Modelo

O problema discutido no presente trabalho apresenta como entrada dados multimodais: vídeos e transcrições de fala. Pensando na faceta do processamento de uma sequência de imagens (frames extraídos de cada clipe), o problema apresenta dois desafios: o processamento de imagens e o processamento de informações sequenciais. Isto posto, uma das formas de abordar este desafio é pensar em um modelo que combine *Convolutional Neural Networks* (CNN) (para o processamento de imagem) e *Recurrent Neural Networks* (RNN) (para o processamento da informação sequencial). Essas particularidades, principalmente o caráter sequencial da informação de entrada, dá pistas de que um modelo baseado em Transformers possa ser um bom caminho a ser seguido.[7]

Transformer é um modelo que surge inicialmente como uma proposta para o Processamento de Linguagem Natural (PLN), mas que cada vez mais vem sendo utilizado em outras áreas relacionadas ao aprendizado de máquinas, inclusive o processamento de imagens. Uma de suas principais características é a sua capacidade de apreender a relação entre os elementos de uma sequência.

Para um primeiro experimento, trabalhou-se somente com o traço de extroversão e propôs-se um modelo baseado em Transformers com saída binária, sendo 1 quando o indivíduo do vídeo de entrada é considerado extrovertido e 0 caso contrário.

O modelo proposto recebe como entrada um clipe curto e o seu respectivo rótulo. Em um primeiro momento, é extraído um número de frames, a ser definido, de cada vídeo. Cada frame é passado por uma instância da rede neural EfficientNet B0 sem a camada de saída. Esta rede, que é pré-treinada, é utilizada como extratora de características dos frames. Por fim, o tensor de características dos frames do vídeo de entrada é submetido a uma análise de componentes principais (PCA) para a redução de dimensionalidade. Ao fim deste processo, tem-se, para cada vídeo de entrada, um tensor da forma (N_FRAMES, PCA_DIM), onde N_FRAMES é o número de frames extraído de cada vídeo e PCA_DIM é a dimensionalidade do vetor de características extraído de cada frame após a sua submissão ao conjunto EfficientNet B0 + PCA.

Esse tensor extraído dos vídeos é, então, utilizado como entrada para uma rede Transformer. A rede utilizada para a classificação é baseada em Paul, 2021[5]: possui uma camada de *positional embedding* seguida de um *transformer encoder*. A camada de *positional embedding* é importante para marcar a informação de ordem nos frames do vídeo, permitindo que a relação temporal possa ser aprendida pelo modelo. O transformer encoder se manteve fiel ao implementado no artigo supracitado.

Para o treinamento, primeiro utilizou-se uma versão reduzida do *dataset*: os vídeos dos primeiro e último quartis referentes à extroversão. Indivíduos presentes em vídeos do primeiro quartil são considerados não extrovertidos (saída igual a 0) e, em vídeos do último quartil, são considerados extrovertidos (saída igual a 1). Adotou-se essa abordagem para a experimentação inicial pois espera-se que a análise dos extremos seja facilitada, permitindo uma avaliação inicial da proposta de modelo.

Com este *dataset* modificado, realizou-se uma série de treinamentos variando diversos hiperparâmetros do modelo e também alterando questões estruturais do modelo. Como hiperparâmetros, foram variados o número de frames extraídos de cada vídeo e a dimensionalidade final do vetor de características de cada imagem. Assim, encontrou-se que a configuração que apresenta os melhores resultados de classificação para o conjunto de dados é N_FRAMES = 20 e PCA_DIM = 2048. Os resultados, para esta configuração, estão apresentados nas Figuras 3, 4 e 5.

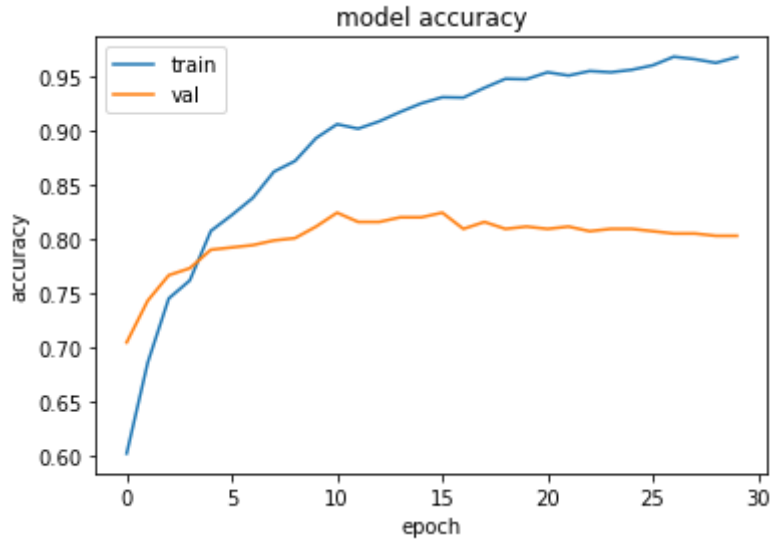

Figura 3. Acurácia do modelo nos conjuntos de treino e de validação conforme época de treinamento.

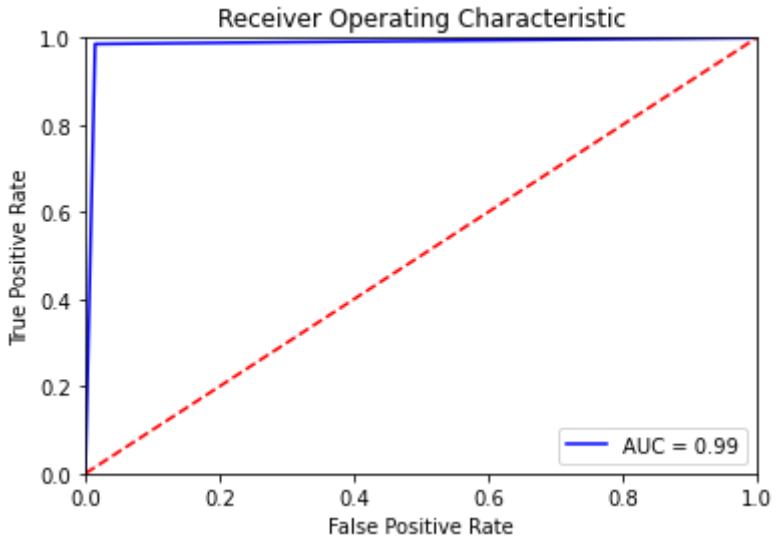

Figura 4. Curva ROC do modelo para o conjunto de treinamento.

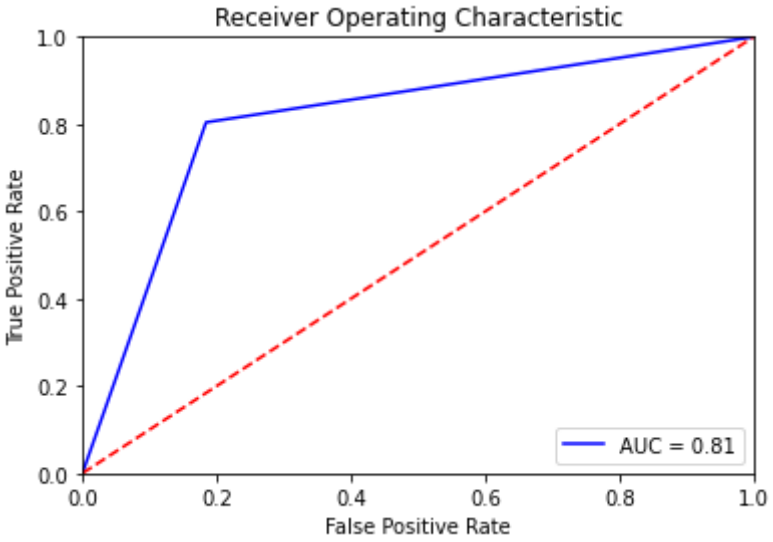

Figura 5. Curva ROC do modelo para o conjunto de validação.

Percebe-se que o modelo parece ser robusto o suficiente para apreender os padrões dos frames e, por fim, realizar a classificação desejada. Entretanto, em todos os testes, o modelo se sobreajustou ao conjunto de treinamento, possuindo uma acurácia reduzida no conjunto de validação. Assim, foram realizados outros testes com o objetivo de lidar com o sobreajuste. Nestes testes, foram experimentadas diversas estratégias de regularização com diversos parâmetros, além de diversas estratégias de *dropout*. Entretanto, não obteve-se melhorias significativas no sobreajuste. Apesar do resultado não tão satisfatório, o modelo parece promissor para lidar com o problema e pretende-se agora treiná-lo no conjunto completo de dados de treinamento; este aumento do número de dados de entrada pode ser um fator importante para lidar com o sobreajuste.

4. Conclusão

O trabalho apresentado trata da predição de primeiras impressões do traço de personalidade de extroversão através de vídeos curtos. Por meio do estudo de *datasets,* foi selecionado o *dataset Chalearn* para estudo. Além disso, conseguiu-se propor um modelo baseado em uma arquitetura ainda não muito explorada para o tema. O modelo foi capaz de predizer de maneira satisfatória se uma pessoa em um vídeo pode ser considerada extrovertida ou não-extrovertida. Apesar do sobreajuste ao conjunto de treinamento, os resultados obtidos indicam que o modelo implementado pode ser adequado para a abordagem do problema.

Por fim, pretende-se iniciar o treinamento do modelo, após as experimentações no conjunto de dados reduzido descrito anteriormente, com todo o conjunto de treinamento do dataset. A tarefa continua sendo a classificação binária extroversão/não extroversão. Para definir a saída de cada vídeo, pretende-se utilizar o *threshold* recomendado pelo artigo do dataset[2]: rótulos acima de 0.5 em extroversão são

considerados extrovertidos (1) e abaixo deste limiar são considerados não extrovertidos (0). Além disso, pretende-se iniciar o treinamento do modelo para a predição dos outros traços de personalidade do *Big five*.

5. Referências

[1] Jacques Junior J. C. S., et al. First Impressions: A Survey on Vision-Based Apparent Personality Trait Analysis. Março de 2022.

[2] Ponce-López V, Chen B, Oliu M, Corneanu C, Clapés A, GuyonI, Baró X, Escalante HJ, Escalera S. (2016) Chalearn lap 2016: first round challenge on first impressions-dataset and results. In: Proceedings of the European conference on computer vision, pp 400–418. Springer

[3] Bradley, R., Terry, M.: Rank analysis of incomplete block designs: The method of paired comparisons. Biometrika 39 (1952) 324–345

[4] Watts, D.J., Strogatz, S.H.: Collective dynamics of small-world networks. Nature 393(6684) (1998) 409–10

[5] Paul, Sayak. Video Classification with Transformers. 2021. Disponível em: https://keras.io/examples/vision/video_transformers/

[6] Escalante, H. J.; Kaya, H.; Salah, A. A.; Escalera, S.; Gucluturk, Y.; Guclu, U.; Baró, X.; Guyon, I.; Jacques Junior, J. C. S.; Madadi, M.; Ayache, S.; Viegas, E.; Gurpinar, F.; Wicaksana, A.S.; Liem, C.C.S.; van Gerven, M. A. J.; van Lier, R. "Modeling, Recognizing, and Explaining Apparent Personality from Videos," IEEE Transactions on Affective Computing (TAC), 2020.

[7] Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems* 30 (2017).

[8] R. R. McCrae, O. P. John. “An introduction to the five-factor model and its applications,” J. Personality, vol. 60, no. 2, pp. 175–215, 1992.

[9] G. J. Boyle, E. Helmes. “Methods of personality assessment” in *The Cambridge Handbook of Personality Psychology*. P. J. Corr & G. Matthews. Eds. Cambridge. U.K.: Cambridge Univ. Press, 2009, pp. 110–126.