Classificação de textos em português utilizando diferentes técnicas de embedding do Word2vec ao GPT

Thomas Edson Cordeiro dos Santos
Programa de Pós-Graduação em Informática - PPGI
Universidade Federal do Espírito Santo - UFES
Vitória, Brasil
thomas.santos@ufes.edu.br

Wharley Borges Ferreira
Programa de Pós-Graduação em Informática - PPGI
Universidade Federal do Espírito Santo - UFES
Vitória, Brasil
wharley.ferreira@edu.ufes.br

***Abstract*—Este artigo apresenta as etapas do trabalho de mapeamento sistemático das técnicas aplicadas para classificação automática de textos em português. Essa revisão tem como objetivo cumprir a atividade 4 da disciplina de Metodologia de Pesquisa e também como requisito para ingresso no Programa de Pós-Graduação em Informática (PPGI) da Universidade Federal do Espírito Santo (UFES). Ao final são apresentadas as técnicas do estado da arte para classificação automática de textos em português e resultados obtidos com o caso particular em estudo.**

***Index Terms*—Processamento de Linguagem Natural, Classificação de textos, português, Embedding**

I. Introdução

O campo de processamento de linguagem natural (do inglês Natural Language Processing - NLP) é uma área da inteligência artificial que tem evoluído rapidamente nos últimos anos. O ChatGPT da OpenAI [1] [2] [3] [4] ganhou as manchetes por apresentar um modelo que consegue ”conversar” com o usuário com uma facilidade e fluência nunca vistos antes. Grandes modelos de linguagem (do inglês Large Language Models - LLM) proprietários apresentam sistematicamente resultados do estado da arte, porém são muito pesados para rodar localmente, semdo disponibilizados por meio de chamadas de API, que possuem custo associado, depende da confiabilidade do serviço e em alguns casos os dados podem ser usados para treinamento de versões posteriores do modelo, levantando questões sobre sigilo e ética. O presente trabalho apresenta a avaliação da aplicabilidade de tecnologias para classificação de textos em português em uma base de dados de ordem de manutenção no sistema SAP. Atualmente essa análise é realizada manualmente por diversas pessoas, com uma cobertura em parte das ordens que são geradas a cada mês. A proposta é que essa análise seja realizada automaticamente em todas as ordens e apenas as ordens mais prováveis de oportunidades de melhorias sejam avaliadas posteriormente por humanos. Nessa abordagem, seria realizada uma cobertura maior das ordens e o uso do trabalho humano seria otimizado. Essas ordens contêm informações possivelmente sensíveis, dessa forma, os dados devem ser tratados em ambientes seguros. Por esse motivo, não foi usada a API pública da OpenAI ou outros serviços. Um dos objetivos da presente revisão é avaliar se a tecnologia por trás do ChatGPT ou as usadas por seus concorrentes treinados especificamente em português do Brasil [5] [6] podem ser aplicadas para classificar com eficiência textos escritos nesse idioma. Foi realizada uma revisão sistemática sobre a classificação de textos em português para verificar quais as técnicas mais promissoras. Ao final do trabalho são apresentados os desempenhos obtidos com diferentes técnicas e é observado que técnicas simples, abertas e consolidadas geram resultados competitivos com as tecnologias mais novas e proprietárias.

II. Ferramentas de embedding de texto utilizadas

Para realizar a classificação de textos é necessário primeiro transformar esses textos em números. Essa tarefa é chamada de embedding e, ao contrário do que pode parecer, não é trivial. As palavras e sentenças devem ser distribuídas em um espaço multidimensional de forma que sentenças similares possuam localizações similares. Foi realizado um mapeamento sistemático com vistas a localizar ferramentas de embedding utilizadas em trabalhos científicos. Um requisito essencial para a seleção das publicações foi buscar as voltadas ao idioma português, preferencialmente o Brasileiro. Considerando a grande revolução que houve nos últimos anos na área de NLP devido ao advento da arquitetura de transformers [1], foram pesquisados artigos publicados a partir de 2018.

A. TF-IDF

De forma a evitar um viés pelo uso exclusivo de técnicas novas, foi utilizada como referência a codificação de textos por meio de Term Frequency-Inverse Document Frequency (TF-IDF), após o que foi realizada a redução de dimensionalidade para 1000 por meio de SVD.

B. Petrolês

Um dos principais trabalhos localizados [6] corresponde ao resultado de uma colaboração interinstitucional liderada pelo Centro de Pesquisas e Desenvolvimento da Petrobras

(CENPES), em parceria com PUC-Rio, UFRGS e PUC-RS, e visa incentivar pesquisas nas áreas de Processamento de Linguagem Natural e Linguística Computacional aplicadas ao domínio de Óleo e Gás (O&G). O Petrolês (https://petroles.puc-rio.ai/) é um repositório de artefatos de Processamento de Linguagem Natural especializados no domínio de petróleo em Português, e tem como objetivo servir como uma referência para os grupos de pesquisas em inteligência artificial e empresas atuantes nesse domínio. Dentre os artefatos disponíveis livremente no repositório foi utilizado no presente trabalho um modelo de word embedding do tipo Word2vec treinado unicamente a partir de dados públicos relacionados ao domínio de O&G (Boletins Técnicos da Petrobras; Teses e Dissertações em assuntos relacionados à indústria de Petróleo; Notas e estudos técnicos da ANP). Para o presente trabalho foi utilizado especificamente o modelo de 300 dimensões PetroVEC_OeG_Word2vec_300d uma vez que, em comparações preliminares entre as diferentes ferramentas disponibilizadas nesse repositório, este mostrou resultados mais promissores.

C. BERTimbau

Uma ferramenta citada recorrentemente nas publicações foi o BERT (do inglês Bidirectional Encoder Representations for Transformers) da Google, em particular uma versão que passou por um fine-tunning em português, o BERTimbau [5]. Esse modelo foi obtido através da biblioteca transformers do Hugging Face distribuída como padrão no pacote ANACONDA. O nome do modelo na biblioteca é o 'neuralmind/bert-base-portuguese-cased'. Como saída, o BERT gera um tensor contendo um vetor de 768 dimensões para cada token. De forma a obter um embedding para toda a sentença, foi realizada a média entre todos os os vetores do tensor, sendo gerada dessa forma apenas um vetor de 768 dimensões para a sentença.

D. SBERT

Também foi utilizada uma versão modificada do BERT específica para o embedding de sentenças, o SBERT [**?**] com uma versão treinada em mais de 50 idiomas [**?**]. Como saída esse modelo gera apenas um vetor de 768 dimensões para cada sentença. O modelo específico utilizado nas simulações foi o "distiluse-base-multilingual-cased-v2" disponível na biblioteca sentence_transformers.

E. GPT

Uma ferramenta de processamento de linguagem natural que ganhou muita atenção recentemente foi o ChatGPT da OpenAI. Em conjunto com os chatbots, a OpenAI disponibiliza as engines por trás desses serviços através de APIs. A PETROBRAS, disponibilizou, em parceria com a Microsoft um serviço com algumas das mais poderosas e recentes ferramentas da OpenAI. No caso do trabalho em questão, não foi cogitado usar a API de chat devido ao limite do tamanho de entrada. O conjunto de dados possui milhares de textos com dezenas de milhares de tokens. Foi utilizada, no entanto uma API de embedding da segunda geração, que de acordo com a própria OpenAI supera as APIs da primeira geração na maioria das tarefas. No presente trabalho foi utilizada especificamente uma versão interna da engine "text-embedding-ada-002". O serviço recebe uma sentença de até 8191 tokens e devolve um embedding com 1536 dimensões. Existe um trabalho bastante detalhado descrevendo o funcionamento da primeira geração de engines de embedding [**?**], porém a documentação relativa à segunda geração de embeddings é mais escassa, o que reflete uma tendência recente dos produtos oferecidos pela OpenAI.

III. Resultado e discussão

Os resultados obtidos (Figura 1) mostram que é possível obter desempenho competitivo com técnicas consolidadas, simples e que rodam localmente, eliminando os custos e a dependência de um serviço fornecido através de uma API [**?**]. A técnica usada como referência foi de pré-processamento de texto (conversão para minúsculo, eliminação de linhas com texto em branco, tokenização, remoção de stopwords, stemming) após o qual foi criada uma matriz do tipo TF-IDF (do inglês Term Frequency–Inverse Document Frequency) e realizada uma redução de dimensionalidade com SVD (do inglês Singular Value Decomposition). Finalmente esses valores foram utilizados para treinar um classificador do tipo Random Forest, com 100 árvores de decisão. Foi disponibilizada na PETROBRAS, em parceria com a Microsoft, um serviço interno com a mesma tecnologia do ChatGPT, o ChatPETROBRAS. Junto com o serviço de chat, foi disponibilizada via API o serviço de "text-embedding-ada-002". De acordo com a documentação disponível no site da OpenAI, esse é o único modelo da segunda geração de embedding, que apresenta desempenho superior aos modelos da primeira geração [**?**] em todas as tarefas, com exceção da tarefa de classificação. O embedding para cada um dos textos foi obtido através dessa API, posteriormente esses vetores foram usados, junto com os rótulos, para treinar um classificador similar ao utilizado com os dados oriundos do TF-IDF. Também foram obtidos os embeddings para os mesmos textos através do BERTimbau [5], uma versão do BERT da Google fine-tunned para o idioma português brasileiro. A conversão foi realizada localmente com o modelo "neuralmind/bert-base-portuguese-cased" disponível através da biblioteca "transformers" na linguagem python. O processo completo levou cerca de 15 h para rodar em um computador com processador i7-12800H 2.40 GHz e 32 GB de memória RAM. Ao contrário do modelo da OpenAI, esse modelo gera um embedding por token. Antes de fornecer a codificação para um classificador similar ao utilizado anteriormente, foi feita a média de todos os embeddings de cada texto, resultando em um embedding da sentença. Um dos motivos de ter sido escolhida a métrica de AUC ROC é que essa métrica é agnóstica ao limiar de decisão selecionado. Como a base de dados é desbalanceada (85/15), é esperado que o modelo seja enviesado a favor da classe predominante.

É possível observar dos resultados que ambas as técnicas (Figura 1) aplicadas conseguem se distanciar bastante de um classificador aleatório (linha tracejada preta) porém ainda estão distantes de um classificador perfeito. Ordens de manutenção

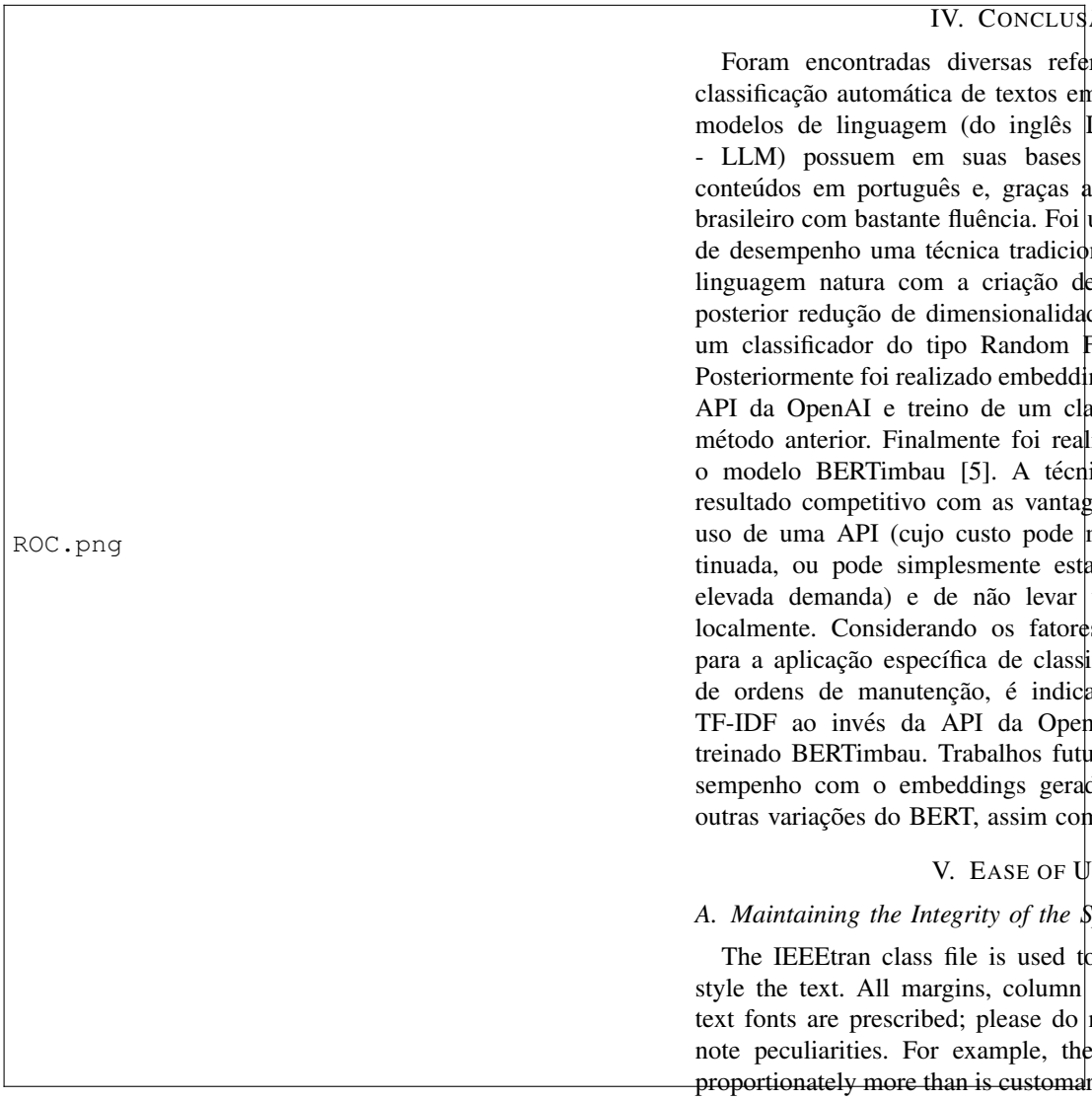

Fig. 1. Curva ROC de classificadores treinados com dados obtidos dos textos longos através de duas técnicas diferentes

TABLE I
TABELA COMPARATIVA DE DIFERENTES TÉCNICAS DE CODIFICAÇÃO

Técnica	Dimensionalidade	AUC-ROC	Camadas	Parâmetros
TF-IDF	1328	0,800	-	-
GPT	1536	0,784	-	-
BERT	768 * n	0,769	12	110 M

são objetos que possuem, além do texto longo associado, diversas outras características relacionais (data de criação, data de encerramento, status, tipo, centro, entre outras). É provável que o uso da classificação por texto longo em conjunto com a classificação por essas características relacionais gere resultados ainda melhores.

IV. Conclusão

Foram encontradas diversas referências voltadas para a classificação automática de textos em português e os grandes modelos de linguagem (do inglês Large Language Models - LLM) possuem em suas bases de treinamento muitos conteúdos em português e, graças a isso, ”falam” o idioma brasileiro com bastante fluência. Foi utilizada como referência de desempenho uma técnica tradicional de processamento de linguagem natura com a criação de uma matriz TF-IDF e posterior redução de dimensionalidade com SVD e treino de um classificador do tipo Random Forest com 100 árvores. Posteriormente foi realizado embedding de sentenças com uma API da OpenAI e treino de um classificador similar ao do método anterior. Finalmente foi realizado o embedding com o modelo BERTimbau [5]. A técnica do TF-IDF mostrou resultado competitivo com as vantagens de não depender do uso de uma API (cujo custo pode mudar, pode ser descontinuada, ou pode simplesmente estar indisponível devido à elevada demanda) e de não levar várias horas para rodar localmente. Considerando os fatores citados anteriormente, para a aplicação específica de classificação de textos longos de ordens de manutenção, é indicada a técnica de matriz TF-IDF ao invés da API da OpenAI ou do modelo pré-treinado BERTimbau. Trabalhos futuros podem avaliar o desempenho com o embeddings gerados pelo SBERT [**?**] ou outras variações do BERT, assim como outros classificadores.

V. Ease of Use

A. Maintaining the Integrity of the Specifications

The IEEEtran class file is used to format your paper and style the text. All margins, column widths, line spaces, and text fonts are prescribed; please do not alter them. You may note peculiarities. For example, the head margin measures proportionately more than is customary. This measurement and others are deliberate, using specifications that anticipate your paper as one part of the entire proceedings, and not as an independent document. Please do not revise any of the current designations.

VI. Prepare Your Paper Before Styling

Before you begin to format your paper, first write and save the content as a separate text file. Complete all content and organizational editing before formatting. Please note sections VI-A–VI-E below for more information on proofreading, spelling and grammar.

Keep your text and graphic files separate until after the text has been formatted and styled. Do not number text heads—LaTeX will do that for you.

A. Abbreviations and Acronyms

Define abbreviations and acronyms the first time they are used in the text, even after they have been defined in the abstract. Abbreviations such as IEEE, SI, MKS, CGS, ac, dc, and rms do not have to be defined. Do not use abbreviations in the title or heads unless they are unavoidable.

B. Units

- Use either SI (MKS) or CGS as primary units. (SI units are encouraged.) English units may be used as secondary units (in parentheses). An exception would be the use of English units as identifiers in trade, such as "3.5-inch disk drive".
- Avoid combining SI and CGS units, such as current in amperes and magnetic field in oersteds. This often leads to confusion because equations do not balance dimensionally. If you must use mixed units, clearly state the units for each quantity that you use in an equation.
- Do not mix complete spellings and abbreviations of units: "Wb/m^2" or "webers per square meter", not "webers/m^2". Spell out units when they appear in text: ". . . a few henries", not ". . . a few H".
- Use a zero before decimal points: "0.25", not ".25". Use "cm^3", not "cc".)

C. Equations

Number equations consecutively. To make your equations more compact, you may use the solidus (/), the exp function, or appropriate exponents. Italicize Roman symbols for quantities and variables, but not Greek symbols. Use a long dash rather than a hyphen for a minus sign. Punctuate equations with commas or periods when they are part of a sentence, as in:

$$a + b = \gamma \tag{1}$$

Be sure that the symbols in your equation have been defined before or immediately following the equation. Use "(1)", not "Eq. (1)" or "equation (1)", except at the beginning of a sentence: "Equation (1) is . . ."

D. LaTeX-Specific Advice

Please use "soft" (e.g., `\eqref{Eq}`) cross references instead of "hard" references (e.g., `(1)`). That will make it possible to combine sections, add equations, or change the order of figures or citations without having to go through the file line by line.

Please don't use the `{eqnarray}` equation environment. Use `{align}` or `{IEEEeqnarray}` instead. The `{eqnarray}` environment leaves unsightly spaces around relation symbols.

Please note that the `{subequations}` environment in LaTeX will increment the main equation counter even when there are no equation numbers displayed. If you forget that, you might write an article in which the equation numbers skip from (17) to (20), causing the copy editors to wonder if you've discovered a new method of counting.

BibTeX does not work by magic. It doesn't get the bibliographic data from thin air but from .bib files. If you use BibTeX to produce a bibliography you must send the .bib files.

LaTeX can't read your mind. If you assign the same label to a subsubsection and a table, you might find that Table I has been cross referenced as Table IV-B3.

LaTeX does not have precognitive abilities. If you put a `\label` command before the command that updates the counter it's supposed to be using, the label will pick up the last counter to be cross referenced instead. In particular, a `\label` command should not go before the caption of a figure or a table.

Do not use `\nonumber` inside the `{array}` environment. It will not stop equation numbers inside `{array}` (there won't be any anyway) and it might stop a wanted equation number in the surrounding equation.

E. Some Common Mistakes

- The word "data" is plural, not singular.
- The subscript for the permeability of vacuum μ_0, and other common scientific constants, is zero with subscript formatting, not a lowercase letter "o".
- In American English, commas, semicolons, periods, question and exclamation marks are located within quotation marks only when a complete thought or name is cited, such as a title or full quotation. When quotation marks are used, instead of a bold or italic typeface, to highlight a word or phrase, punctuation should appear outside of the quotation marks. A parenthetical phrase or statement at the end of a sentence is punctuated outside of the closing parenthesis (like this). (A parenthetical sentence is punctuated within the parentheses.)
- A graph within a graph is an "inset", not an "insert". The word alternatively is preferred to the word "alternately" (unless you really mean something that alternates).
- Do not use the word "essentially" to mean "approximately" or "effectively".
- In your paper title, if the words "that uses" can accurately replace the word "using", capitalize the "u"; if not, keep using lower-cased.
- Be aware of the different meanings of the homophones "affect" and "effect", "complement" and "compliment", "discreet" and "discrete", "principal" and "principle".
- Do not confuse "imply" and "infer".
- The prefix "non" is not a word; it should be joined to the word it modifies, usually without a hyphen.
- There is no period after the "et" in the Latin abbreviation "et al.".
- The abbreviation "i.e." means "that is", and the abbreviation "e.g." means "for example".

An excellent style manual for science writers is [7].

F. Authors and Affiliations

The class file is designed for, but not limited to, six authors. A minimum of one author is required for all conference articles. Author names should be listed starting from left to right and then moving down to the next line. This is the author sequence that will be used in future citations and by indexing services. Names should not be listed in columns nor group by affiliation. Please keep your affiliations as succinct as possible (for example, do not differentiate among departments of the same organization).

G. Identify the Headings

Headings, or heads, are organizational devices that guide the reader through your paper. There are two types: component heads and text heads.

Component heads identify the different components of your paper and are not topically subordinate to each other. Examples include Acknowledgments and References and, for these, the correct style to use is "Heading 5". Use "figure caption" for your Figure captions, and "table head" for your table title. Run-in heads, such as "Abstract", will require you to apply a style (in this case, italic) in addition to the style provided by the drop down menu to differentiate the head from the text.

Text heads organize the topics on a relational, hierarchical basis. For example, the paper title is the primary text head because all subsequent material relates and elaborates on this one topic. If there are two or more sub-topics, the next level head (uppercase Roman numerals) should be used and, conversely, if there are not at least two sub-topics, then no subheads should be introduced.

H. Figures and Tables

a) Positioning Figures and Tables: Place figures and tables at the top and bottom of columns. Avoid placing them in the middle of columns. Large figures and tables may span across both columns. Figure captions should be below the figures; table heads should appear above the tables. Insert figures and tables after they are cited in the text. Use the abbreviation "Fig. 2", even at the beginning of a sentence.

TABLE II
TABLE TYPE STYLES

Table Head	**Table Column Head**		
	Table column subhead	***Subhead***	***Subhead***
copy	More table copy[a]		

[a]Sample of a Table footnote.

Fig. 2. Example of a figure caption.

Figure Labels: Use 8 point Times New Roman for Figure labels. Use words rather than symbols or abbreviations when writing Figure axis labels to avoid confusing the reader. As an example, write the quantity "Magnetization", or "Magnetization, M", not just "M". If including units in the label, present them within parentheses. Do not label axes only with units. In the example, write "Magnetization (A/m)" or "Magnetization {A[m(1)]}", not just "A/m". Do not label axes with a ratio of quantities and units. For example, write "Temperature (K)", not "Temperature/K".

ACKNOWLEDGMENT

The preferred spelling of the word "acknowledgment" in America is without an "e" after the "g". Avoid the stilted expression "one of us (R. B. G.) thanks ...". Instead, try "R. B. G. thanks...". Put sponsor acknowledgments in the unnumbered footnote on the first page.

REFERENCES

Please number citations consecutively within brackets [1]. The sentence punctuation follows the bracket [2]. Refer simply to the reference number, as in [3]—do not use "Ref. [3]" or "reference [3]" except at the beginning of a sentence: "Reference [3] was the first ..."

Number footnotes separately in superscripts. Place the actual footnote at the bottom of the column in which it was cited. Do not put footnotes in the abstract or reference list. Use letters for table footnotes.

Unless there are six authors or more give all authors' names; do not use "et al.". Papers that have not been published, even if they have been submitted for publication, should be cited as "unpublished" [4]. Papers that have been accepted for publication should be cited as "in press" [5]. Capitalize only the first word in a paper title, except for proper nouns and element symbols.

For papers published in translation journals, please give the English citation first, followed by the original foreign-language citation [6].

REFERENCES

[1] G. Eason, B. Noble, and I. N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955.

[2] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[3] I. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[4] K. Elissa, "Title of paper if known," unpublished.

[5] R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.

[6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[7] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

00

REFERENCES

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper$_files/paper/2017/file/3f5ee243547dee91fbd0$ $Paper.pdf$

[2] A. Radford and K. Narasimhan, "Improving language understanding by generative pre-training," 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:49313245

[3] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," 2019. [Online]. Available: https://www.semanticscholar.org/paper/Language-Models-are-Unsupervised-Multitask-Learners-Radford-Wu/9405cc0d6169988371b2755e573cc28650d14dfe

[4] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," 2020.

[5] F. Souza, R. Nogueira, and R. Lotufo, "BERTimbau: pretrained BERT models for Brazilian Portuguese," in *9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23 (to appear)*, 2020.

[6] D. d. S. M. Gomes, F. C. Cordeiro, B. S. Consoli, N. L. Santos, V. P. Moreira, R. Vieira, S. Moraes, and A. G. Evsukoff, "Portuguese word embeddings for the oil and gas industry: Development and evaluation," *Computers in Industry*, vol. 124, 2021, cited by: 12. [Online]. Available: