Classificação automática de textos de *User-Generated Content* utilizando Aprendizagem de Máquina Supervisionado

Iolanda Victoria Morais Ramos [1], Jackson Wilke da Cruz Souza [1,2]

[1]Instituto de Ciência, Tecnologia e Inovação - Universidade Federal da Bahia (UFBA), Camaçari/BA - Brasil

[2]Programa de Pós-Graduação em Língua e Cultura - Universidade Federal da Bahia (UFBA), Salvador/BA - Brasil

iolanda.ramos@ufba.br, jackcruzsouza@gmail.com

Abstract. *This study aims to develop an automatic text classifier for User-Generated Content from the DANTE-Stocks corpus. The classification algorithm was trained in a supervised manner, using labels provided by human annotators and subsequently associated with various vectorization methods. In the end, a classifier was generated that performs very close to human-level performance in identifying the three proposed classes, namely: (i) well-structured, (ii) moderately structured, and (iii) poorly structured.*

Resumo: *Este estudo visa a construção de um classificador automático de textos User-Generated Content do corpus DANTE-Stocks. O algoritmo de classificação foi treinado de forma supervisionada, utilizando rótulos propostos por anotadores humanos e, posteriormente, associado a diferentes métodos de vetorização. Ao final, gerou-se um classificador que performa bastante próximo ao desempenho humano, ao identificar três classes propostas de organização dos tweets, a saber: (i) bem, (ii) mediamente e (iii) mal estruturado.*

1. Introdução

As redes sociais têm desempenhado um papel crucial para a produção, circulação e recepção de conteúdos de interesses para a sociedade. Com a expansão das redes sociais, os usuários assumiram um papel cada vez mais ativo como geradores de conteúdo. Os conteúdos gerados por usuários (em inglês, *User-Generated Content* (UGC), segundo Wyrwoll (2014), podem significar uma grande contribuição para o desenvolvimento e progresso intelectual da sociedade.

Para a área de Processamento de Línguas Naturais (PLN), UGC apresenta desafios únicos de processamento dadas suas características ligadas à linguagem e ao modo de circulação de mensagens. Nesse sentido, o conteúdo gerado pode não seguir padrões linguísticos e estruturais ligados à norma culta da língua, apresentando grande diversidade nessas questões.

Para lidar com esses desafios, é necessário um conjunto de técnicas e recursos em PLN, sobretudo para o Português do Brasil (PB), língua ainda em fase de desenvolvimento de recursos para o processamento de textos de UGC. Técnicas como, por exemplo, classificação e agrupamentos desses textos são de grande importância para aprimorar a *qualidade* de identificação de padrões e fenômenos linguísticos, e as *dimensionalidades* quanto à gerenciabilidade da performance dos modelos, facilitando, posteriormente, análises e modelagem linguístico-computacionais.

Partindo da classificação prévia de Pereira e Souza (2024), construímos um classificador automático para *tweets*/postagens do *corpus* DANTEStock [Di Felippo *et al.* 2021]. Foram testadas duas abordagens de vetorização dos dados linguísticos para a construção dos algoritmos de Aprendizado de Máquina (AM) supervisionado a partir do paradigma teórico conexionista/neural [Monard e Baranauskas 2003], o qual busca simular o processamento de informações inspirado no modelo biológico do sistema nervoso. Além disso, as classificações foram submetidas a avaliações *quantitativas* (a partir de métricas clássicas de AM) e *qualitativas* (avaliação humana). Como resultado, foi promovida a classificação do *corpus* em função da estrutura de cada *tweet*/postagem.

Destaca-se que o emprego de diferentes técnicas de avaliação e desenvolvimento de ferramentas e recursos em PLN devem ser compreendidos como uma ponte fundamental entre linguística e computação. Rodrigues, Souza e Santos (2022) destacam que essa interação é "de mão dupla". Isso significa que, por um lado, a linguística é essencial para desenvolver e melhorar os recursos que as máquinas usam para interpretar a linguagem. Por outro lado, as ferramentas computacionais também podem ajudar a refinar e validar os conhecimentos linguísticos criados pelos humanos.

Para tanto, este artigo está organizado em cinco seções, além desta Introdução. Na Seção 2, apresentamos a metodologia utilizada nesta pesquisa. Na Seção 3, destacamos os resultados no processo de treinamento do modelo supervisionado desenvolvido para a tarefa de classificação de textos de UGC. Por fim, na Seção 4, tecemos considerações finais e indicações de trabalhos futuros.

2. Metodologia

Neste trabalho, propusemos a criação de um modelo de AM supervisionado para classificação do *corpus* DANTEStock [Di Felippo *et al.* 2021]. Tal *corpus* é constituído por *tweets* ligados ao cenário de ações da bolsa de valores do Brasil. O *corpus* é composto por 4,518 *tweets* e seus identificadores únicos, que foram compilados a partir da coleta automática de postagens do X/Twitter, em 2014.

Para a criação e o treinamento de um classificador neste trabalho, foram testados algoritmos de diferentes paradigmas, sendo o algoritmo *Multilayer classifier* - MLP [Haykin 1994] o escolhido por apresentar melhor desempenho em termos de métricas quantitativas e avaliação qualitativa. Destaca-se que o ambiente de desenvolvimento foi o Colaboratory do Google. Ademais, as bibliotecas utilizadas em Python foram extraídas do *scikit-learn* [Kramer 2016].

Para a tarefa de treinamento do modelo, uma amostra menor do *corpus* contendo 180 *tweets* foi rotulada por três anotadores, como apontado por Pereira e Souza (2024). Neste trabalho, os autores propuseram três classes considerando a organização sintática, semântica e estrutural das sentenças, a saber: *bem estruturado*, com 81 exemplares*, mediamente estruturado,* com 59, e *mal estruturado*, com 39. Os algoritmos foram treinados observando o texto dos *tweets* e tendo como alvo de predição as classes propostas pelos anotadores, como exemplificado em (1), retirado de Pereira e Souza (2024).

(1)
a) *Bem estruturado:* Ano passado eu falei que até o final de 2104 #PETR4 estaria abaixo de R$10,00 mas acho que errei, não vai demorar tanto.

b) *Mediamente estruturado:* vai, oibr4. um troux... ops... investidor precisa pagar as minhas férias
c) *Mal estruturado:* &lt,Alexander Cruz3 *-*

O *tweet* (1a) foi classificado como *bem estruturado*, pois sua estrutura não prejudica em nada a compreensão do conteúdo, mesmo sem um contexto informacional maior. Já o *tweet* (1b) foi classificado como *mediamente estruturado*, já que possui compreensão limitada dada a sua (des)organização sintática e semântica. Por fim, o *tweet* (1c), classificado como *mal estruturado*, tem uma baixa compreensão em aspecto semântico e falta estruturação lógica que permita a compreensão da mensagem.

Para lidar com o desbalanceamento dos dados, o modelo foi desenvolvido considerando as técnicas de validação cruzada [Netto e Maciel 2021] para melhorar o aprendizado dos critérios de cada classe e a vetorização das instâncias com base no modelo pré-treinado de Bertimbau [Souza e Nogueira 2020].

Após a etapa de treinamento do modelo, passamos para a etapa de análise quanti e qualitativa dos resultados de desempenho de cada modelo. Nessa etapa, foram avaliadas quantitativamente as métricas de desempenho [Netto e Maciel 2021], com base em Precisão (P), Revocação (R), Medida-F (MF) e Acurácia (A). Quanto à avaliação qualitativa, selecionamos de forma aleatória um exemplo de *tweet*/postagem e analisamos se a classificação proposta pelo modelo fazia sentido se comparada com a avaliação humana da estrutura sentencial do exemplo.

3. Resultados e discussão

De forma preliminar, cada modelo classificador foi avaliado de acordo com suas métricas utilizadas neste trabalho (P, R, MF e A). O classificador foi treinado utilizando duas diferentes técnicas de vetorização, a saber *Term Frequency - Inverse Data Frequency - TF-IDF* [Moreira 2024] e *Bidirectional Encoder Representations from Transformers (BERT),* mais especificamente usando a variação treinada para o PB, o *BERTimbau*. Em ambos os casos, o modelo MLP se mostrou mais adequado. A Tabela 1 resume o resultado dos desempenhos obtidos ao classificar nossa amostra.

Classes	**Classificador / Medidas**							
	MLP - TF-IDF				**MLP - BERT**			
	P	**R**	**MF**	**A**	**P**	**R**	**MF**	**A**
Bem estruturado	0.58	0.72	0.64	**0.57**	0.77	0.80	0.78	**0.74**
Mediamente estruturado	0.47	0.35	0.40	**0.57**	0.68	0.65	0.67	**0.74**
Mal estruturado	0.75	0.67	0.71	**0.57**	0.78	0.78	0.78	**0.74**

Tabela 1. Métricas obtidas de cada modelo em etapa inicial

O desempenho do modelo MLP pode, possivelmente, ser explicado pelo próprio paradigma conexionista que o modelo possui, sendo capaz de absorver melhor as nuances mais complexas de representações propostas na etapa de aprendizado baseada nos rótulos fornecidos. Outro aspecto importante a ser considerado é a melhora significativa de desempenho do modelo quando associado ao método de vetorização utilizando BERT. Em termos de acurácia, os modelos saíram de 57% para 74%; já em termos de MF, a classe que mais bem foi beneficiada com a abordagem foi a de textos *mediamente estruturados*, saindo de 40% para 67%. Essa melhora pode possivelmente ser explicada

pela capacidade do BERT de entender o contexto das palavras, o que é crucial para a classificação de *tweets*, em que o contexto pode alterar fortemente o significado. Além disso, o BERTimbau é um modelo não apenas treinando para o PB, mas também treinado com conteúdo proveniente de redes sociais. Isso faz com que o classificador lide melhor com as nuances semânticas captadas pela vetorização.

Além disso, submetemos o classificador a novos exemplos de *tweets*/postagens retirados para avaliarmos a capacidade de generalização, como exemplificado em (2).

(2)

a) Bem estruturado: Vamos ver se mesmo em dia fraco, ganha alguma força ...RT @Live_Trade: #ecor3 fechando 15' acima 12,53 já fica interessante"
b) *Mediamente estruturado:* que linda era esa mina chabonnn
c) *Mal estruturado:* @victoriabril_ forra contesta

Nos exemplos, a categoria atribuída pelo modelo MLP a (2a) foi *bem estruturada*: a estrutura do *tweet* não prejudica sua compreensão, embora apresente desvios de pontuação, por exemplo. Por outro lado, em (2b) o modelo classificou a instância como *mal estruturada*, o que difere da avaliação humana, que rotulou a mesma instância como *mediamente estruturada*. Apesar dos desvios ortográficos presentes, é possível identificar uma estrutura mínima na mensagem, a qual poderia ser mais bem compreendida se considerada dentro de um contexto adequado. Por fim, (2c) foi classificado como *mal estruturado*: a estrutura do *tweet* não permite nenhuma compreensão acerca da mensagem.

4. Considerações finais

Os objetivos deste trabalho, que incluíam a construção de um classificador automático para *tweets* e postagens do *corpus* DANTEStock, a aplicação de técnicas de AM e a avaliação das classificações, foram devidamente alcançados. Por se tratar de um estudo preliminar, optou-se por trabalhar com uma amostra reduzida do *corpus* original, em virtude do alto custo de recurso humano para a classificação manual dos dados, essencial para garantir a precisão e confiabilidade das informações. Apesar desta questão, os resultados quantitativos e qualitativos demonstraram a capacidade do classificador de lidar com a categorização de textos de UGC. Além disso, o agrupamento do *corpus* com base na estrutura sentencial dos textos tornará possível anotações linguísticas e/ou a identificação de padrões importantes sobre a diversidade e complexidade dos textos.

Para a tarefa de classificação de textos de UGC, o uso de AM supervisionado permite que o modelo aprenda de maneira aproximada ao desempenho humano. Trata-se do levantamento de características importantes em função do rótulo-alvo, melhorando a precisão com que o modelo performa e generaliza em exemplos semelhantes, *a posterior*.

Destaca-se que esta pesquisa contribui de maneira significativa para o projeto POeTiSA: POrtuguese processing - Towards Syntactic Analysis and parsing, que visa desenvolver ferramentas e aplicações linguístico-computacionais para o PB. A integração da análise de UGC em diferentes teorias linguístico-computacionais, pode auxiliar, por exemplo, na identificação de fenômenos ainda não descritos no PB. Para trabalhos futuros, serão explorados outros algoritmos de AM, além de desenvolver estratégias para melhorar o balanceamento dos textos, garantindo a preservação da naturalidade dos dados durante o processo de ajuste e modelagem.

Agradecimentos

Este trabalho foi realizado no âmbito do Centro de Inteligência Artificial da Universidade de São Paulo (C4AI -http://c4ai.inova.usp.br/), com o apoio da Fundação de Amparo à Pesquisa do Estado de São Paulo (processo FAPESP #2019/07665-4) e da IBM. Este projeto também foi apoiado pelo Ministério da Ciência, Tecnologia e Inovações, com recursos da Lei N. 8.248, de 23 de outubro de 1991, no âmbito do PPI-Softex, coordenado pela Softex e publicado como Residência em TIC 13, DOU 01245.010222/2022-44. Este projeto também foi apoiado pela Universidade Federal da Bahia, através do programa de bolsas de iniciação científica - ações afirmativas (PIBIC-AF) 2023/2024.

Referências

Breiman, L. (2001). Random Forests. *Machine Learning*, 45, pp. 5–32. DOI: https://doi.org/10.1023/A:1010933404324

Di Felippo, A. *et al.* (2021). "Descrição Preliminar do Corpus DANTEStocks: Diretrizes de Segmentação para Anotação segundo Universal Dependencies". In: *Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*. Porto Alegre, Brasil: SBC, p. 335-343. DOI: https://doi.org/10.5753/stil.2021.17813

Haykin, S. (1994). *Neural Networks: A Comprehensive Foundation*. Prentice Hall PTR.

Kramer, O., e Kramer, O. (2016) Scikit-learn. *Machine learning for evolution strategies*, p. 45-53

Mann, W. C., e Thompson, S. A. (1988). Rhetorical Structure Theory: Toward a Functional Theory of Text Organization. *Text-interdisciplinary Journal for the Study of Discourse*, 8(3), pp. 243–281. Disponível em: https://www.sfu.ca/rst/05bibliographies/bibs/Mann_Thompson_1988.pdf

Mikolov, T. *et al*. (2013). Efficient Estimation of Word Representations in Vector Space. *Preprint.* Disponível em: http://arXiv:1301.3781

Monard, M. C., E Baranauskas, J. A. (2003). Conceitos sobre Aprendizado de Máquina. *Sistemas Inteligentes: Fundamentos e Aplicações*, 1(1), p. 1.

MOREIRA, V. P. (2024). Recuperação de Informação. In: NUNES, M. G. (Org.), *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*, 2. ed. [s.l.]: BPLN. Disponível em: https://brasileiraspln.com/livro-pln/2a-edicao/parte-aplicacoes/cap-ir/cap-ir.html

NETTO, A., e MACIEL, F. (2021). *Python para Data Science e Machine Learning Descomplicado*. Rio de Janeiro: Editora Alta Books, 397p.

Pereira, M.A., Souza, J.W.C. (2024). Subsídios Linguísticos para classificação automática de textos de User-Generated Content. In *Anais do XV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*. Porto Alegre: SBC.

Rodrigues, R., Souza, J. W. C., e Santos, R. L. S. (2022). "Descrição Linguística e Aprendizado de Máquina: Análise de Verbos Locativos do Espanhol". *Cadernos de Estudos Linguísticos*, Campinas, SP, 64(00), p. e022038. DOI: https://doi.org/10.20396/cel.v64i00.8666995

Souza, F., Nogueira, R., E Lotufo, R. (2020). BERTimbau: Pretrained BERT Models for Brazilian Portuguese. In: Cerri, R., and Prati, R. C. (Eds.), *Intelligent Systems. BRACIS 2020. Lecture Notes in Computer Science*, vol. 12319, Springer, Cham. DOI: https://doi.org/10.1007/978-3-030-61377-8_28

Wyrwoll, C. (2014). *User-Generated Content. Social Media,.*In C. Wyrwoll (Ed.), Social Media: Fundamentals, Models, and Ranking of User-Generated Content. Springer Fachmedien, p *11–45.* DOI: https://doi.org/10.1007/978-3-658-06984-1_2

Zhang, H. (2004). The Optimality of Naive Bayes. In: *Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference*. Florida/USA: American Association for Artificial Intelligence. p.1-6.