Subsídios Linguísticos para classificação automática de textos de *User-Generated Content*

Mateus Araújo Pereira[1], Jackson Wilke da Cruz Souza [1]

[1] Instituto de Letras (ILUFBA) – Universidade Federal da Bahia (UFBA) – Salvador/BA

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

pereiramateus@ufba.br, jackcruzsouza@gmail.com

Abstract: *This study aims to classify the structures of User-Generated Content (UGC) texts using the DANTE-stocks corpus, which consists of tweets about the financial market. The texts were manually analyzed and classified based on semantic, cohesive, and coherence criteria according to their linguistic structure, resulting in three classes: (i) well-structured, (ii) moderately structured, and (iii) poorly structured. The integration of these approaches provides a foundation for developing applications in the field of Natural Language Processing related to UGC texts.*

Resumo: *Este estudo visa classificar as estruturas de textos User-Generated Content (UGC), usando o corpus DANTE-stocks, composto por tweets sobre o mercado financeiro. Os textos foram analisados e classificados manualmente com critérios semânticos, coesivos e de coerência em função da estrutura linguística, resultando em três classes: (i) bem, (ii) mediamente e (iii) mal estruturado. A integração dessas abordagens oferece embasamento para o desenvolvimento de aplicações no âmbito do Processamento de Linguagem Natural com relação a textos de UGC.*

1. Introdução

Krumm, Davies e Narayanaswami (2008) definem *User-Generated Content* (UGC), como quaisquer conteúdos criados por usuário em uma plataforma *online*, integrando um conjunto de formatos, como textos, fotos, vídeos, comentários em fóruns ou redes sociais. Essa definição enfatiza o ambiente participativo e colaborativo da UGC, destacando conteúdos, em sua espontaneidade, produzidos por usuários devido à permissividade do ambiente e do gênero textual em que esses textos são produzidos. A UGC é caracterizada por uma variedade de fenômenos linguísticos, como o uso de abreviações, neologismo e estruturas não convencionais, como corrobora Di-Felippo *et al.*, (2021).

A maior parte das pesquisas em Processamento de Linguagem Natural (PLN), até o momento, tem se concentrado na análise de textos formais, em detrimento de textos de UGC. Em virtude da diversidade e informalidade presentes nesses textos, exigem de abordagens direcionadas a compreender suas características únicas. Para que seja possível realizar análises linguísticas considerando particularidades do gênero textual e da linguagem em textos de UGC, objetivamos neste trabalho propor classes que levem em conta aspectos semânticos e estruturais. Nesse sentido, serão

apresentados os resultados do estudo piloto de classificação do *corpus* [Di-Felippo *et al.* 2021]. O estudo teve foco na identificação e categorização dos *tweets* em três classes, a saber: *bem estruturado, mediamente estruturado e mal estruturado.* Ademais, detalharemos a metodologia aplicada, as abordagens linguísticas adotadas durante o processo, e os resultados preliminares alcançados.

Para tanto, utilizou-se como conjunto de dados o *corpus* DANTE-stocks [DiFelippo *et al.*, 2021], que é composto por aproximadamente 6,700 *posts/tweets* extraídos da plataforma X/Twitter, pertencentes ao domínio do mercado financeiro. Esses textos são caracterizados por uma linguagem curta e direta, contendo jargões e observações específicas ao contexto de uso e suportes midiáticos e abordam informações sobre ações, investimentos, notícias econômicas e análises de mercado das bolsas de valores.

Este trabalho está organizado em 3 seções, além desta introdução e da consideração final. Na seção 2, apresentamos uma visão sobre a literatura com estudos que abordam a análise da linguagem em textos de UGC, destacando seus fenômenos linguísticos específicos do gênero textual. Na seção 3, detalhamos a metodologia utilizada na classificação dos *tweets* e quais critérios foram considerados nesse processo. Na seção 4, exploramos as duas abordagens linguísticas aplicadas à classificação dos *tweets* e sua contribuição para a determinação dessas classes. Por fim, na consideração final, discutimos a importância de uma reflexão linguística que vá além das normas cultas da língua e propomos direções para trabalhos futuros.

2. Trabalhos relacionados

Diversos estudos abordam a análise da linguagem em textos de UGC, na perspectiva da caracterização de fenômenos linguísticos específicos, como abreviações, neologismos e estruturas não convencionais. Tagg (2012) investigou a linguagem em *SMS* e redes sociais, observando como a espontaneidade e a informalidade influenciam a estrutura e o conteúdo comunicacional. A autora argumenta que os usuários adaptam os textos às redes sociais, utilizando abreviações e *emojis*, por exemplo, para transmitir mensagens de maneira rápida e demonstrar emoção.

No mesmo contexto, Androutsopoulos (2011) analisou a variedade linguística e a inovação lexical em interações *online*, evidenciando a dinâmica criativa da linguagem digital. O autor pontua que as redes sociais fornecem um espaço para experimentação linguística, permitindo aos usuários criarem vocabulários e formarem novas expressões. Diante disso, é crucial expor a importância de considerar o contexto comunicativo na análise linguística de textos de UGC, que desafiam as normas da GT.

Além desses estudos, Eisenstein (2013) aplicou técnicas de PLN para analisar grande quantidade de textos de UGC, concentrando-se na identificação de padrões linguísticos e fonológicos presentes nesses textos. Eisenstein destaca que a escrita em redes sociais pode refletir o modo de pronúncia das palavras, com a substituição de letras para representar sons específicos. O autor argumenta que a escrita, nesse contexto, tem aspectos fonológicos, ao demonstrar uma conexão entre a fala e a escrita.

É importante salientar que essas pesquisas demonstram como abordagens computacionais podem ser usadas para perceber a diversidade e a singularidade da linguagem em ambientes digitais. Dessa forma, os resultados dos trabalhos contribuem para o desenvolvimento de modelos de PLN para lidar com a variedade linguística presente em textos de UGC.

Somado a isso, estudos anteriores [Tagg 2012; Eisenstein 2013], focam nos fenômenos linguísticos como abreviações e a inovação lexical nas redes sociais, mas poucos se dedicam a classificar esses textos quanto a sua estrutura. Assim, esta pesquisa é um dos poucos estudos a propor uma categorização de *tweets* baseando-se em critérios de coesão, coerência e semântica. Trabalhos mais recentes [Longaretti 2021; Steinhauser e Botassini 2023] indicam a ocorrência de fenômenos linguísticos voltados aos níveis semântico, gramatical, lexical, discursivo e prosódico em redes sociais.

3. Metodologia

Este estudo passou por duas etapas metodológicas. Na primeira, foi conduzido um estudo exploratório nos *tweets*, com ênfase nos aspectos morfossintáticos - a análise da estrutura e organização das palavras dentro das frases, como combinações de artigo + substantivo + verbo. Para tal, foram analisadas 70 postagens selecionadas aleatoriamente do *corpus* DANTE-stocks [Di Felippo *et al.* 2021], com o objetivo de identificar padrões estruturais característicos. Apesar disso, os resultados não evidenciaram regularidade, chegando-se a 70 possibilidades de combinações. Nesse contexto, na segunda etapa definiram-se três classes para os *tweets*, baseadas em aspectos semânticos (compreensão da mensagem), coesão (conexão entre elementos linguísticos do texto) e coerência (organização lógica dos itens linguísticos).

Dada a inviabilidade de propor um padrão morfossintático para cada *tweet* do *corpus*, foram propostas três classes que consideram a estrutura linguística de cada um deles. Para tanto, baseamo-nos em critérios de naturezas semântica, coesiva e coerente para cada uma das classes propostas, como demonstrado no Quadro 1.

CRITÉRIOS	**CLASSES**		
	Bem estruturado	**Mediamente estruturado**	**Mal estruturado**
Semântica Compreensão textual	Postagem totalmente compreensível	Postagem parcialmente compreensível, mas requer certo conhecimento do domínio	Postagem totalmente incompreensível.
Coesão Conexão entre proposições	Boa conexão.	Conexão imprecisa, apesar de presente	Pouquíssima ou nenhuma conexão
Coerência Organização lógica	Boa organização lógica	Organização lógica imprecisa	Carência e/ou ausência de organização lógica

Quadro 1. Descrição dos critérios entre as classes.

No Quadro 1 descrevem-se os critérios usados para categorizar os *tweets* nas três classes propostas. *Tweets bem estruturados* têm alta compreensão semântica e coesiva, mesmo sem conhecimento prévio do contexto ou domínio. *Tweets mediamente estruturados* têm semântica parcialmente compreensível e coesão limitada, necessitando de algum conhecimento prévio. Por fim, *tweets mal estruturados* têm baixa compreensão semântica, conexões fracas e falta de lógica, dificultando a interpretação do conteúdo.

A classificação dos *tweets* foi realizada manualmente por três anotadores, sem que tivessem acesso à anotação individual do grupo. Ao fim desse processo os resultados foram comparados automaticamente. Para tanto, quando havia concordância total sobre a classificação entre os três anotadores, o *tweet* era rotulado diretamente com a classe indicada. Porém, quando houve casos de discordância, utilizou-se da regra da

maioria: a classe com maior voto entre os anotadores predominava como a escolha final para a rotulação do *tweet.*

4. Resultados e Discussões

Como resultado, foram analisados 180 *tweets*, distribuídos entre as classes *bem estruturada* (82 tweets), *mediamente estruturada* (59) e *mal estruturada* (39). Essa etapa envolveu a leitura cuidadosa de cada *tweet*, considerando os aspectos semânticos e discursivos para essa classificação. Após a determinação dos critérios de cada uma das classes, os *tweets* foram classificados manualmente, como demonstrado no Quadro 2.

TWEETS	**CLASSIFICAÇÃO**
1) PETR4 subiu na bolsa 13,50. Muito bem, estou surpreso com o resultado.	Bem estruturado
2) 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.	Bem estruturado
3) vai, oibr4. um troux... ops... investidor precisa pagar as minhas férias 4) que linda era esa mina chabonnn	Mediamente estruturado
5) &lt,Alexander Cruz3 *-* 6) @victoriabril_ forra contesta	Mal estruturado

Quadro 2. Exemplos de *tweets* e sua classificação.

No Quadro 2, o *tweet* em (5), classificado como *Mal estruturado*, não permite compreensão sobre o conteúdo, pois há apenas a indicação de um nome e a inclusão de um *emoji* (“ *-* ”), tido como um recurso paralinguístico, desempenhando uma função discursiva. Já o conteúdo de (1), classificado como *Bem estruturado*, é compreensível por si só, se enquadrando nos três critérios: a semântica é direta e acessível, a coesão é mantida entre os elementos linguísticos, e a coerência permite uma organização lógica que contribui o entendimento do texto como um todo. Por fim, (3), classificado como *mediamente estruturado*, apresenta uma sequência lógica do conteúdo, mas que apresenta elementos que podem prejudicar a estruturação do conteúdo, como a interrupção da fala (em “um troux…”) e a correção do que foi dito (indicado por “ops…”).

5. Considerações finais

Neste estudo procuramos delimitar critérios para classificar *tweets* em função da estrutura de composição e compreensão do conteúdo. Esses métodos poderão ser utilizados para o treinamento e teste de algoritmos de Aprendizado de Máquina para gerar classificadores automáticos, permitindo escalabilizar o processo. Proporcionando uma compreensão sobre os aspectos comunicativos em plataformas digitais. A integração de abordagens linguísticas e computacionais na análise de *tweets*, representa um avanço na capacidade de compreender e modelar a comunicação digital, oferecendo novas ferramentas para a análise de grandes volumes de dados textuais e melhorando a interação entre humanos e máquinas em ambientes de redes sociais.

É importante destacar a limitação de perspectivas prescritivistas e normativas, alinhadas à Gramática Tradicional nesse tipo de pesquisa. Eventualmente, *tweets* que foram previamente classificados como *mediamente estruturados* poderiam ser classificados como *bem estruturados*, como a postagem (2) do Quadro 1, por exemplo. Assim, em análises futuras será imprescindível considerar outras perspectivas

linguísticas que levem em conta o uso da língua, como a Gramática Funcional, compreendendo limites e contribuições metodológicas e teóricas mútuas.

Por fim, destaca-se a importância deste trabalho no escopo de anotação de *corpus*. O resultado deste trabalho permitirá identificar quais *tweets* possuem diferentes características estruturais e linguísticas que poderão ser úteis para a identificação de fenômenos discursivos, por exemplo. Isso permitirá, portanto, a aplicação de modelos teóricos nesse sentido, avançando o estado da arte em PLN para o PB.

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. Além disso agradecemos ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), a Fundação de Amparo à Pesquisa do Estado da Bahia (FAPESB) pelo financiamento e suporte.

Referências

Androutsopoulos, J. (2011). Language Change and Digital Media: A Review of Conceptions and Evidence. *Standard languages and language standards in a changing Europe*, 1, pp.145-159.

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

Eisenstein, J. (2013). Phonological Factors in Social Media Writing. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing* (EMNLP), (pp.11-19). Atlanta: Association for Computational Linguistics.

Krumm, J.; Davies, N. e Narayanaswami, C. (2018) User-generated content. *IEEE Pervasive Computing*, v. 7, n. 4, pp.10-11.

Longaretti, R. B. (2021). *O difícil de dizer em texto de instrução ao sósia sobre o trabalho docente: uma análise de fenômenos linguísticos prosódicos*. 163f. Dissertação (Mestrado em Letras) – Programa de Pós-Graduação em Letras – Universidade Tecnológica Federal do Paraná, Pato Branco, Paraná, 2021.

Steinhauser, V. L. F., e Botassini, J. O. M. (2023). Vem sabonetar aqui fora! Um estudo multissistêmico do verbo Sabonetar sob influência do twitter e dos reality shows. *Papéis: Revista do Programa de Pós-Graduação em Estudos de Linguagens - UFMS*, *27*(53), pp.114-138.

Tagg, C. (2012). *Discourse of Text Messaging*: Analysis of SMS Communication. London: Continuum.