Universidade Estadual de Maringá
Centro de Ciências Exatas
Departamento de Física

Trabalho de Conclusão de curso

Análise de Padrões Textuais em Vestibulares com Modelos de Linguagem e Métodos de Sistemas Complexos

Acadêmico: Lucas Augusto Lima de Souza

Orientador: Prof. Dr. Haroldo Valentin Ribeiro

Maringá, 10 de Janeiro de 2025.

Universidade Estadual de Maringá
Centro de Ciências Exatas
Departamento de Física

Trabalho de Conclusão de curso

Análise de Padrões Textuais em Vestibulares com Modelos de Linguagem e Métodos de Sistemas Complexos

Monografia apresentada ao Departamento de Física da Universidade Estadual de Maringá, sob orientação do professor Dr. Haroldo Valentin Ribeiro, como parte dos requisitos para obtenção do título de licenciado em Física.

Acadêmico: Lucas Augusto Lima de Souza

Orientador: Prof. Dr. Haroldo Valentin Ribeiro

Maringá, 10 de Janeiro de 2025.

Lucas Augusto Lima de Souza

Análise de Padrões Textuais em Vestibulares com Modelos de Linguagem e Métodos de Sistemas Complexos

Monografia apresentada ao Departamento de Física da Universidade Estadual de Maringá, sob orientação do professor Dr. Haroldo Valentin Ribeiro, como parte dos requisitos para obtenção do título de licenciado em Física.

Banca Examinadora

Orientador - Prof. Dr. Haroldo Valentin Ribeiro
Departamento de Física - Universidade Estadual de Maringá

Prof. Dr. Sergio de Picoli Junior
Departamento de Física - Universidade Estadual de Maringá

Prof. Dr. Ervin Kaminski Lenzi
Departamento de Física - Universidade Estadual de Maringá

Maringá-PR
2025

AGRADECIMENTOS

Agradeço aos professores do curso que contribuíram direta ou indiretamente para a construção deste trabalho, com suas aulas e orientações. Agradeço especialmente ao Prof. Dr. Haroldo Valentin Ribeiro, pois, sem ele, este trabalho não teria saído do papel. Aos meus amigos de graduação, que foram tão resilientes em conseguirem terminar o curso, e aos colegas que, apesar de terem saído, guardo um carinho enorme. Aos amigos da vida, que estiveram comigo durante essa etapa, minha gratidão por todo apoio e companhia. Agradeço também à minha namorada Geovanna, que sempre esteve ao meu lado desde o início, oferecendo apoio incondicional e palavras de incentivo nos momentos mais desafiadores. Sua paciência e motivação foram fundamentais. Por fim, minha eterna gratidão à minha família: minha mãe, Sonia, meu pai, Roberto, e minha irmã, Byanca, que serviu de inspiração nesta jornada acadêmica. Pelo apoio, incentivo e compreensão ao longo de toda essa jornada, vocês foram fundamentais para que eu chegasse até aqui.

EPÍGRAFE

"A felicidade é a pretensão ilusória de converter um instante de alegria em eternidade."

(Clóvis de Barros)

RESUMO

Este trabalho analisa os textos de provas de vestibulares da Universidade Estadual de Maringá (UEM) utilizando técnicas de sistemas complexos e processamento de linguagem natural (PLN). Por meio do modelo BERT, investigaram-se padrões textuais associados a respostas corretas e incorretas, além de diferenças estruturais nas questões. A metodologia envolveu a coleta, limpeza e organização dos dados, seguidas por análises quantitativas, como a Lei de Zipf e o mapeamento vetorial dos textos das questões. Embora padrões claros não tenham sido identificados para separar alternativas corretas e incorretas, o modelo demonstrou boa precisão na classificação de questões por matérias (77%) e grandes áreas (91%). Os resultados indicam que os textos dos vestibulares seguem padrões linguísticos naturais, sem viés sistemático que favoreça a identificação de respostas corretas, destacando a justiça do processo seletivo. Conclui-se que o uso de técnicas avançadas de PLN podem ampliar a compreensão dos padrões textuais em avaliações educacionais e contribuir para o aprimoramento de ferramentas pedagógicas.

Palavras-chave: Processamento de Linguagem Natural, Sistemas Complexos, Lei de Zipf.

ABSTRACT

This study analyzes the texts from university entrance exams of the Universidade Estadual de Maringá (UEM) using complex systems techniques and natural language processing (NLP). Through a BERT model, textual patterns associated with correct and incorrect answers, as well as structural differences in the questions, were investigated. The methodology included the collection, cleaning, and organization of data, followed by quantitative analyses, such as Zipf's Law and text embedding of the questions. Although no clear patterns were identified to distinguish correct from incorrect alternatives, the model demonstrated good accuracy in classifying questions by subjects (77%) and broad areas (91%). The results indicate that the exam texts follow natural linguistic patterns without systematic bias that would favor the identification of correct answers, emphasizing the fairness of the selection process. It is concluded that the use of advanced NLP techniques can enhance the understanding of textual patterns in educational assessments and contribute to the improvement of pedagogical tools.

Keywords: Natural Language Processing, Complex Systems, Zipf's Law.

LISTA DE ABREVIATURAS E SIGLAS

BERT	Bidirectional Encoder Representations from Transformers
ELMo	Embeddings from Language Models
LLM	Large Language Models
LM	Modelagem linguística
LSTM	Long Short Term Memory
NLM	Neural Language Models
PAS	Processo de Avaliação Seriada
PLM	Pre-trained language models
PLN	Processamento de linguagem natural
SLM	Small Language Models
UEM	Universidade Estadual de Maringá

SUMÁRIO

1. INTRODUÇÃO

A escrita é uma das maiores invenções da humanidade e tem desempenhado um papel essencial na preservação do conhecimento e no avanço das civilizações ao longo dos séculos, muitas sociedades a consideram até como um presente dos deuses **(1,2)**. Desde os primeiros registros cuneiformes na antiga Mesopotâmia até os sofisticados sistemas de escrita digital de hoje, a capacidade de registrar e transmitir informações de maneira eficiente transformou o modo como a sociedade interage e compartilha conhecimento **(2)**. Johannes Gutenberg, no século XV, foi responsável pela invenção da prensa tipográfica permitindo a produção em massa de textos e a disseminação de ideias em larga escala. Esse avanço tecnológico permitiu que a revolução Científica acontecesse, por exemplo **(3)**. Com o tempo, a escrita evoluiu e tornou-se cada vez mais acessível, com inovações como a máquina de escrever no século XIX e, posteriormente, o computador no século XX. Hoje, vivemos na era digital, em que a criação e o compartilhamento de textos ocorrem de maneira instantânea e em uma quantidade massiva, plataformas online, redes sociais e bancos de dados digitais produzem diariamente volumes gigantescos de informação textual e por isso a era digital, com suas infinitas possibilidades de armazenamento e compartilhamento, ampliou a diversidade linguística e aumentou a complexidade dos textos produzidos, criando novos desafios e oportunidades para a análise dessas informações **(4)**.

No artigo de Oliver **(5),** o autor discute a correspondência entre os escritores Mário de Andrade, um dos principais representantes do Modernismo no Brasil, e Manuel Bandeira, que tinha uma inclinação por modelos literários clássicos e europeus. O estudo revela as dificuldades de Andrade em codificar a língua falada, que é naturalmente fluida, e seu desejo de descobrir a essência da chamada "brasilidade" em um país marcado pela imigração. A correspondência entre os dois autores debate sobre suas visões distintas da linguagem através da literatura no contexto do Modernismo.

No contexto educacional, particularmente em sistemas de avaliação como os vestibulares, o estudo de textos e questões se torna cada vez mais relevante. Provas de vestibulares são instrumentos centrais na avaliação do conhecimento dos estudantes, abrangendo áreas diversas do saber e exigindo habilidades variadas. Esses exames não são apenas uma ferramenta para medir conhecimento, mas também revelam padrões textuais complexos que podem influenciar tanto na formulação das perguntas quanto no desempenho dos candidatos. Cada questão traz em si um conjunto de informações que deve ser compreendido, interpretado e analisado pelos alunos, e as diferenças na estrutura dessas

perguntas podem impactar a maneira como as respostas são elaboradas. Mário e Manuel já tinham um embate a respeito da linguagem falado no país e como eles gostariam de se comunicar, seguindo essa mesma lógica é justo estudarmos se a linguagem usada em textos de vestibulares também segue uma tendência ou até mesmo uma preferência sentimental oculta que permeia as avaliações.

Este trabalho, propõe-se a analisar provas de vestibulares através de técnicas de sistemas complexos e do processamento de linguagem natural (PLN). Em particular, com foco no uso do modelo *BERT (Bidirectional Encoder Representations from Transformers)*, que utiliza de aprendizado profundo para o tratamento de textos sendo capaz de capturar nuances linguísticas e contextuais, possibilitando a classificação de questões e respostas em padrões. Este estudo busca identificar características textuais que podem estar associadas às respostas verdadeiras ou falsas, analisando se há uma relação entre a estrutura das perguntas e a natureza das respostas. Além disso, será investigada a aplicação de técnicas quantitativas para explorar padrões textuais e diferenças entre respostas verdadeiras e falsas, utilizando métodos que permitem identificar diferenças na formulação e no número de palavras entre elas.

Assim, este trabalho pretende contribuir para uma melhor compreensão dos padrões textuais presentes em provas de vestibulares, com o objetivo de aprimorar tanto a formulação de exames quanto às estratégias de ensino e avaliação. Ao empregar modelos de linguagem e técnicas de sistemas complexos, o estudo também visa demonstrar o potencial dessas ferramentas no campo educacional, abrindo novas possibilidades para a análise de textos em grandes volumes e para a compreensão das dinâmicas envolvidas nos processos de avaliação.

2. REVISÃO DE LITERATURA

A evolução da escrita também gerou uma incrível diversidade de idiomas e palavras, refletindo a riqueza e a complexidade das culturas humanas, já que cada idioma possui seu próprio conjunto de regras e expressões, contribuindo para um vasto mosaico de comunicação global. Apesar dessa quantidade diversa de idiomas, regras e culturas, é observado que a comunicação segue um certo padrão, qualquer que seja o meio estudado. Em 1935, George Zipf notou que, ao ordenar palavras pela frequência com que apareciam em diversos textos, a palavra mais frequente tendia a ser usada cerca de duas vezes mais do que a segunda palavra mais comum, aproximadamente três vezes mais do que a terceira, e essa relação se repetia ao longo do que estava sendo analisado **(6)**. Uma das possíveis explicações dadas por Zipf foi baseada no conceito de que tanto o falante quanto o ouvinte desejam minimizar o esforço durante a comunicação, porém não é possível concluir isso de forma precisa. Usando argumentos estatísticos, Zipf definiu que a distribuição de frequências *f(r)* seria descrita por sua lei,

$$f(r) \sim r^{-\alpha}, \tag{1}$$

denotando r o ranque de frequência de uma palavra ($r = 1$ representa a palavra mais frequente, $r = 2$ indica a segunda e assim por diante). A Lei de Zipf (Equação 1) é, portanto, uma lei de potência com α sendo o seu expoente, geralmente próximo de 1 em diversos textos e idiomas.

2.1 REDES NEURAIS

Apesar de a utilização de métodos para análise de texto não ser uma novidade na área de sistemas complexos, o interesse nesse tipo de trabalho vem crescendo significativamente com o advento dos chamados "*large language models"* (LLMs). Esses modelos são programas treinados com conjuntos de dados que visam reconhecer e gerar textos. Seu funcionamento se baseia no aprendizado de máquina (*machine learning*) e no aprendizado profundo (*deep learning*), utilizando redes neurais **(7)**.

As redes neurais artificiais são técnicas computacionais que utilizam um modelo matemático inspirado na estrutura neural de organismos inteligentes, adquirindo conhecimento por meio da experiência. Compostas por unidades chamadas neurônios ou vértices, essas redes podem ter centenas ou milhares de neurônios organizados em camadas.

Cada neurônio recebe entradas, processa essas informações através de funções de ativação e gera saídas **(8)**.

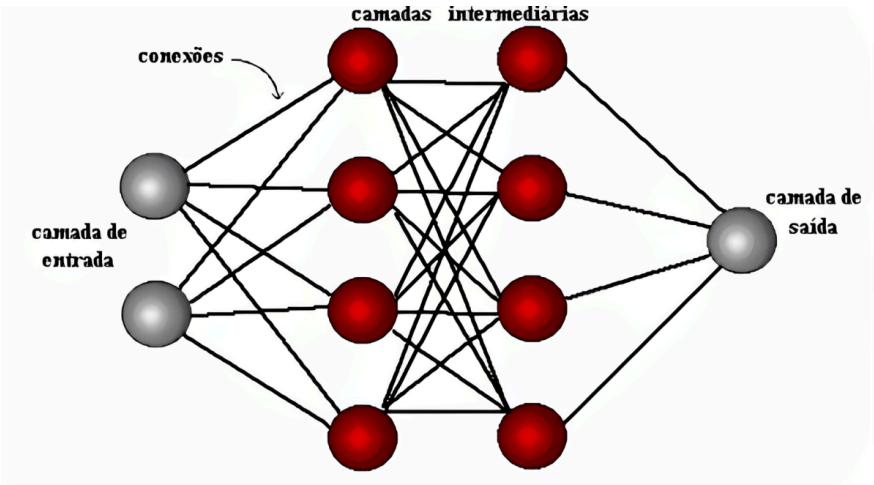


Figura 1. Organização em camadas de redes neurais **(9)**.

A organização em camadas, ilustrada na Figura 1, permite que as redes neurais sejam eficazes em capturar representações complexas dos dados, especialmente em PLN que é utilizado neste trabalho. Elas funcionam através de três abordagens principais: aprendizado supervisionado, onde recebem dados rotulados; aprendizado não supervisionado, que não conta com um agente externo para guiar as respostas; e aprendizado por reforço, onde um avaliador externo fornece feedback sobre as respostas da rede **(9)**. No caso do aprendizado supervisionado, as unidades de processamento, geralmente conectadas por canais de comunicação associados a pesos específicos, realizam operações apenas com os dados locais que recebem por meio de suas conexões.

O que chamamos de comportamento inteligente de uma rede neural artificial vem das interações entre as unidades de processamento. Seu modelo de operação foi proposto por McCulloch e Pitts em 1943 **(10)** e pode ser resumida da seguinte maneira:

- Apresentações de sinais na entrada;
- Cada sinal é multiplicado por um número, ou peso, que indica a sua influência na saída da unidade;
- A soma ponderada é feita para produzir um nível de atividade
- Caso o nível de atividade exceda um certo limite (*threshold*) a unidade produz uma determinada resposta de saída.

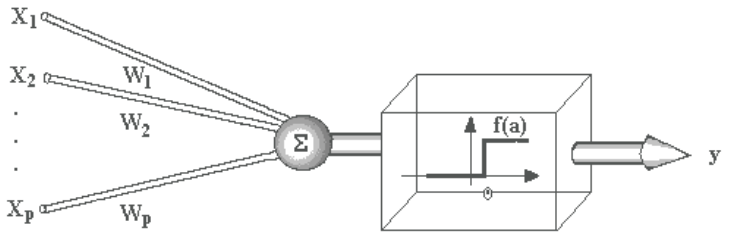


Figura 2. Esquema de unidade McCulloch - Pitts **(9)**.

Suponha que tenhamos p sinais de entrada $X_1, X_2, ..., X_p$ e pesos $w_1, w_2, ..., w_p$ e limitador t (Figura 2); com sinais assumindo valores booleanos (0 ou 1) e pesos valores reais. Neste modelo, o nível de atividade a é dado por:

$$a = w_1X_1 + w_2X_2 + ... + w_pX_p$$

A saída y é dada por

$$y = 1, \text{ se } a >= t$$

ou

$$y = 0, \text{ se } a < t.$$

Anos depois, desenvolvido nas décadas de 1970 e 1980, o algoritmo de *backpropagation* é um método de treinamento amplamente utilizado em redes neurais multicamadas. Ele permite que a rede aprenda ajustando os pesos das conexões com base no erro da saída, que é calculado ao comparar a saída prevista com a saída real. Esse erro é, então, propagado para trás através da rede, possibilitando que cada neurônio ajuste seus pesos de maneira a minimizar essa discrepância. O ajuste dos pesos é realizado utilizando o gradiente do erro em relação aos pesos, uma abordagem que viabiliza a otimização do modelo **(11)**.

O treinamento desse tipo de programa (*Large Language Models* - LLMs) tem como objetivo entender como os caracteres, palavras e frases funcionam juntos, pois o aprendizado profundo envolve a análise probabilística de dados não estruturados, o que acaba permitindo que o modelo reconheça distinções entre partes do conteúdo **(12)**. Um dos modelos que melhor exemplifica isso é o sistema da OpenAI - ChatGPT (*Generative Pre-trained Transformer*) **(13)** que atualmente é amplamente utilizado por diversos tipos de pessoas com variados objetivos.

A modelagem linguística (*Language Modeling* - LM) é uma das principais abordagens utilizadas para melhorar a inteligência linguística das máquinas, envolvendo em geral, modelagem da probabilidade de sequências de palavras para prever a probabilidade do futuro **(12)**. O desenvolvimento da modelagem linguística pode ser dividido em quatro fases, começando com a fase das modelagens estatísticas (*Small Language Model* - SLM) cujo exemplo clássico é o modelo *n-gram*, que analisa sequências contínuas de *n* itens em um texto ou fala, fazendo a predição do próximo item na sequência se baseando em um modelo de Markov, onde a probabilidade de transição para um estado futuro depende apenas do estado atual, sem considerar os estados anteriores. Em outras palavras, dado o estado atual do sistema, o próximo estado é independente do passado, dependendo apenas da situação presente **(14)**. Na segunda fase do desenvolvimento da modelagem linguística, surgiram os Modelos Neurais de Linguagem (*Neural Language Models* - NLMs), também conhecidos como modelagem neural da linguagem, introduzindo o uso de redes neurais para prever a probabilidade da próxima palavra em uma sequência, levando em consideração as palavras anteriores na sequência. A terceira fase é marcada por modelos que capturam o contexto e o significado das palavras em uma frase ou texto, conhecidos como modelos linguísticos pré-treinados (*Pre-trained language models* - PLMs), esses modelos empregam redes neurais para aprender representações vetoriais das palavras, levando em consideração o contexto em que cada palavra ocorre. São exemplos os modelos: Word2vec, ELMo e BERT **(15)**. A quarta fase é marcada pelos *Large Language Models* (LLMs) já citadas anteriormente.

2.2 TIPOS DE PLMs

2.2.1 Word2Vec

O Word2Vec é uma técnica que se destaca por sua acessibilidade e simplicidade, características que estão diretamente ligadas à sua popularidade. A facilidade de implementação e os bons resultados em tarefas de análise semântica, permitem que até pesquisadores com pouca experiência em Processamento de Linguagem Natural (PLN) possam integrá-lo facilmente em seus projetos. Com código aberto e diversos modelos pré-treinados disponíveis, a experimentação é acessível e com ampla possibilidade de adoção global, destaca-se também pela capacidade de capturar relações semânticas entre palavras, como no exemplo clássico “rei está para rainha assim como homem está para mulher”, o que evidencia seu potencial em análises linguísticas **(16)**.

O objetivo principal dessa técnica é criar uma representação numérica para as

palavras **(17)**, fazendo a transformações para vetores multidimensionais, onde essa quantidade de dimensões é definida no momento do treinamento do Word2Vec. Como o vetor simboliza um ponto no hiperespaço, o Word2Vec posiciona cada palavra de um vocabulário em um ponto específico dentro desse espaço. Palavras com significados semelhantes tendem a estar próximas entre si. Para realizar essa representação, uma rede neural é treinada com base em um corpus de texto **(18)**. Para que a representação seja eficaz, esse corpus precisa conter uma grande quantidade de palavras.

2.2.2 ELMo

O ELMo (*Embeddings from Language Models*) utiliza redes neurais LSTM (*Long Short-Term Memory*) bidirecionais para gerar representações contextuais das palavras. Isso significa que ele ajusta o significado de uma palavra de acordo com o contexto em que ela está inserida, permitindo que uma mesma palavra assume diferentes significados dependendo da frase **(19)**. Por exemplo, a palavra *banco* pode se referir a uma instituição financeira ou a um assento, dependendo do contexto, e diferenciar essas variações é especialmente útil para lidar com palavras ambíguas. Essa habilidade se deve ao treinamento com grandes volumes de texto que permite ao modelo aprender a prever palavras e identificar significados diversos.

2.2.3 BERT

Embora o ELMo represente um avanço significativo ao considerar o contexto bidirecional, o BERT (*Bidirectional Encoder Representations from Transformers*) vai além, capturando essas relações de maneira ainda mais precisa através da sua arquitetura baseada em *Transformers*, que processa todas as palavras de uma frase simultaneamente, e consegue encontrar relações complexas entre termos, independentemente de sua posição **(20)**.

Seu treinamento consiste na combinação da técnica de previsão de palavras mascaradas, em que o modelo aprende a prever palavras ocultas em frases, e também na modelagem de relações de sentenças, que ensina o modelo a entender conexões entre diferentes partes de um texto. Essas abordagens permitem ao BERT capturar detalhes e semânticas com maior sofisticação, superando o ELMo em tarefas de PLN. Além de tudo, é possível aumentar a flexibilidade do modelo adaptando-o a tarefas específicas e fornecendo dados adicionais por meio de um processo chamado fine-tuning, que otimiza ainda mais seu desempenho em aplicações práticas.

2.2.4 Diferenças

As diferenças entre os LMs ficam mais claras analisando cada uma lado a lado através das datas de lançamento, recursos e possibilidades oferecidas (Tabela 1).

Modelo	Data de lançamento	Prós	Contras	Diferenças
Word2Vec	2013	- Simplicidade e facilidade de implementação - Rápido e eficiente para tarefas básicas	- Representações estáticas (não contextuais) - Não captura nuances semânticas complexas	- Gera vetores fixos para palavras, independentemente do contexto - Usa uma arquitetura simples (CBOW ou Skip-Gram) em vez de redes neurais profundas como BERT e ELMo
ELMo	Fevereiro de 2018	- Representações contextuais dinâmicas - Integração fácil com modelos existentes	- Menos flexível que BERT - Baseado em LSTMs, o que pode ser menos eficiente que transformadores	- Usa LSTMs bidirecionais em vez de transformadores - Menos complexo no pré-treinamento em comparação com BERT - Representações obtidas a partir de uma única sequência

BERT	Outubro de 2018	- Representações contextuais dinâmicas - Alto desempenho em tarefas de PLN	- Requer mais recursos computacionais - Implementação mais complexa	- Usa transformadores em vez de LSTMs - Treinamento bidirecional desde o início - Pré-treinamento com Modelagem de Linguagem Mascarada e Previsão da Próxima Sentença

Tabela 1. Relação entre data de lançamento, prós, contras e diferenças entre as LMs Word2Vec, ELMo e BERT.

Para que o entendimento fique ainda melhor acerca do funcionamento, podemos utilizar o esquema da Figura 3, que compara a arquitetura dos três modelos.

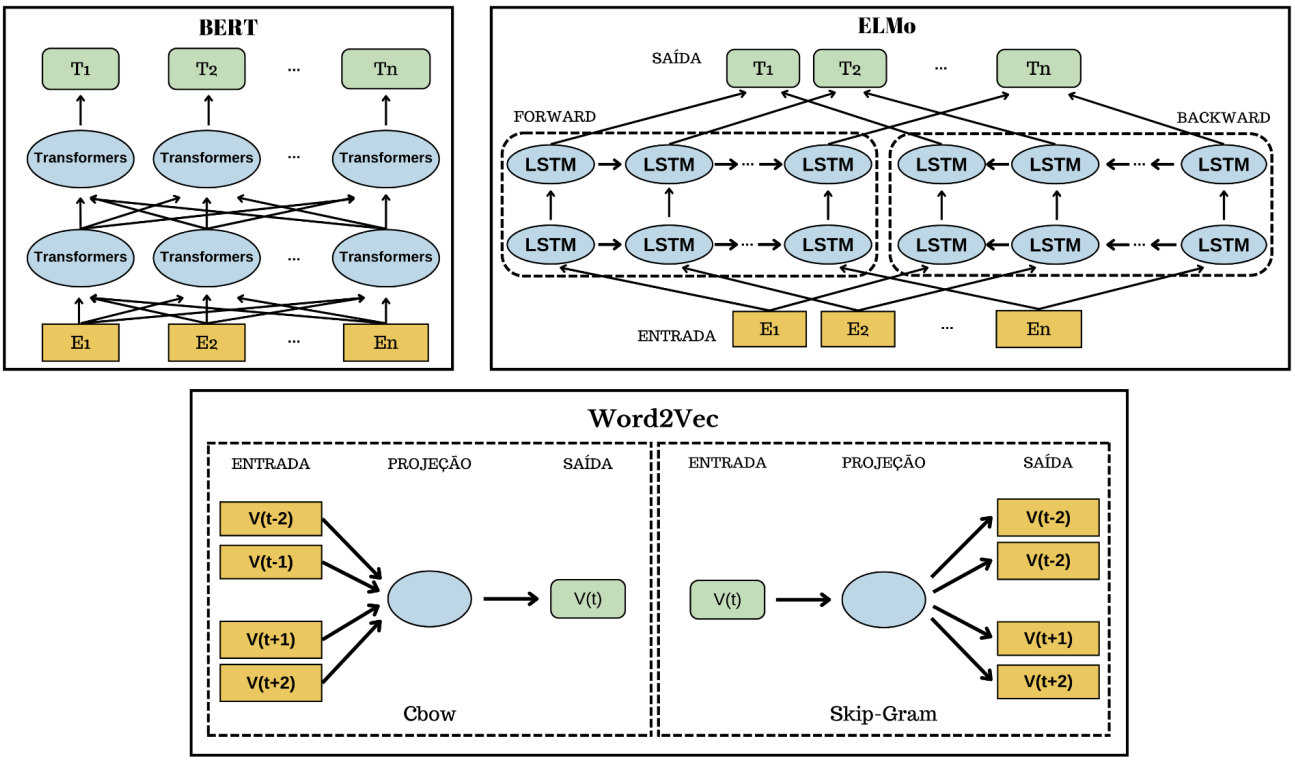


Figura 3. Representação esquemática das LMs. Adaptado de Devlin et al. (2018).

3. METODOLOGIA

Para conduzir a análise dos textos de vestibulares, no contexto dos vestibulares da Universidade Estadual de Maringá (UEM), foi adotado um processo estruturado de limpeza e organização dos dados, seguido por uma análise automatizada utilizando a linguagem Python. Inicialmente, todas as provas foram coletadas e organizadas conforme a data e o tipo de exame, categorizando-os em modalidades como Vestibular de Verão, Vestibular de Inverno e as diferentes etapas do Processo de Avaliação Seriada (PAS – etapas 1, 2 e 3). Como as provas estavam disponíveis apenas em formato PDF no site oficial da UEM, foi solicitado à Comissão do Vestibular Unificado o envio dos arquivos no formato .docx, uma vez que a conversão dos PDFs para texto poderia comprometer a integridade dos dados. Esse procedimento facilitou as etapas subsequentes de organização e limpeza. As provas utilizadas datam de 2015 até 2023 e a análise foi realizada somente sobre a prova de conhecimentos gerais.

A extração do conteúdo textual dos arquivos .docx foi realizada com o uso da biblioteca '*docx2txt*' em Python. Os arquivos foram organizados em diretórios e, a partir da função '*glob*', foi feita uma listagem de arquivos permitindo uma manipulação automatizada dos dados. Na sequência, os textos extraídos passaram por uma etapa de limpeza, onde foram removidos elementos não relevantes para a análise, como folhas de instrução, cabeçalhos, rodapés e espaços para anotação de respostas (Figuras 4 e 5).

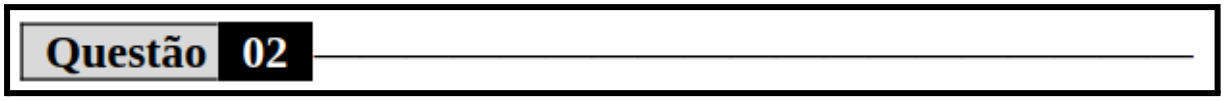


Figura 4. Exemplo de cabeçalho de questão que foi removido.

Figura 5. Exemplo de rodapé contido na prova que foi removido.

Essa limpeza foi realizada utilizando expressões regulares com a biblioteca '*re*', garantindo a padronização do formato das questões e alternativas. Entretanto, devido à presença de equações matemáticas, que em sua maioria estavam representadas como imagens e não em formato textual, essas informações não puderam ser processadas adequadamente, resultando em uma redução na quantidade de dados analisados especificamente em questões

que envolvem matemática.

Após a remoção dos elementos desnecessários, as questões de múltipla escolha foram estruturadas para análise. Cada questão foi organizada em uma lista, com suas respectivas alternativas sendo identificadas por números padronizados ("01", "02", "04", "08", "16"). O código foi desenvolvido para isolar e extrair essas alternativas, assegurando que todos os dados estivessem no formato adequado para a análise. Em seguida, foi implementada uma função chamada '*verifica_numero_itens*', que verificava se o número de alternativas em cada questão estava de acordo com o esperado. Essa verificação foi realizada utilizando vetores e operações matemáticas com a biblioteca '*numpy*', de forma a deixar o processo de contagem e validação mais eficiente.

Para a visualização dos resultados, a biblioteca '*matplotlib*' foi utilizada na produção de gráficos que ilustram a distribuição das alternativas e outros padrões observados nas questões. Essas representações gráficas facilitaram a interpretação dos dados processados. Por fim, através do modelo de linguagem natural BERT, foi feita uma análise dos dados para classificar as questões nas suas respectivas matérias/grandes áreas, e realizar a comparação entre respostas verdadeiras e falsas, permitindo uma visualização clara das diferenças e padrões encontrados. Todos os passos foram executados através da plataforma *Jupyter Notebook*.

4. RESULTADOS E DISCUSSÃO

A análise da Lei de Zipf com base nos textos de vestibulares da Universidade Estadual de Maringá (UEM) confirma que as palavras seguem um padrão típico de frequência linguística. Como esperado, as palavras mais comuns são aquelas de função gramatical, como preposições, artigos e conjunções **(21)**.

A Tabela 2 mostra as 10 palavras mais frequentes e suas respectivas contagens:

Palavra	**Frequência**
de	7541
a	4491
e	3898
o	3708
que	3164
do	2480
da	2373
é	1756
em	1717
um	1326

Tabela 2. Palavras mais frequentes presentes nas questões dos vestibulares e suas respectivas contagens.

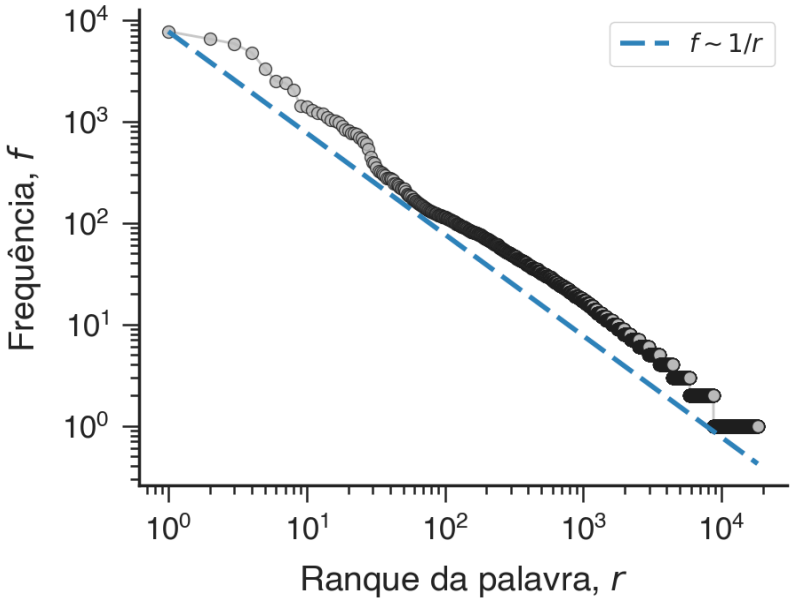


Gráfico 1: Lei de Zipf nos textos de vestibulares da UEM.

Esses resultados refletem a organização da língua portuguesa, em que palavras funcionais possuem alta frequência por serem fundamentais para a estruturação das frases. No caso dos textos analisados, a palavra mais frequente, "de", aparece 7.541 vezes, enquanto "a" e "e" ocupam o segundo e terceiro lugares (Tabela 2), respectivamente.

Palavras mais frequentes, essenciais para a coesão textual, carregam baixo peso semântico e por consequência têm menor impacto na compreensão de ideias mais complexas, essas, aparecem nas regiões de menor densisade no Gráfico 1. Palavras de menor frequência, geralmente compostas por substantivos e verbos mais específicos, carregam maior carga de significado, sendo estas que muitas vezes determinam a dificuldade interpretativa das questões, essas por consequência estão relacionadas na área de maior densidade do gráfico.

A curva resultante segue uma distribuição linear negativa representando a aderência dos dados à Lei de Zipf, isso reforça a universalidade desse padrão da linguagem natural, indicando que os textos de vestibular seguem as mesmas características estatísticas de outras produções linguísticas.

Partindo para uma análise mais geral, pensando nas quantidades de questões presentes em todas as provas da UEM, é possível ver que existe uma ocorrência maior nas questões de ciências da natureza (Biologia, Química, Física) seguidas por Português, Matemática, linguagens e ciências humanas (Gráfico 2).

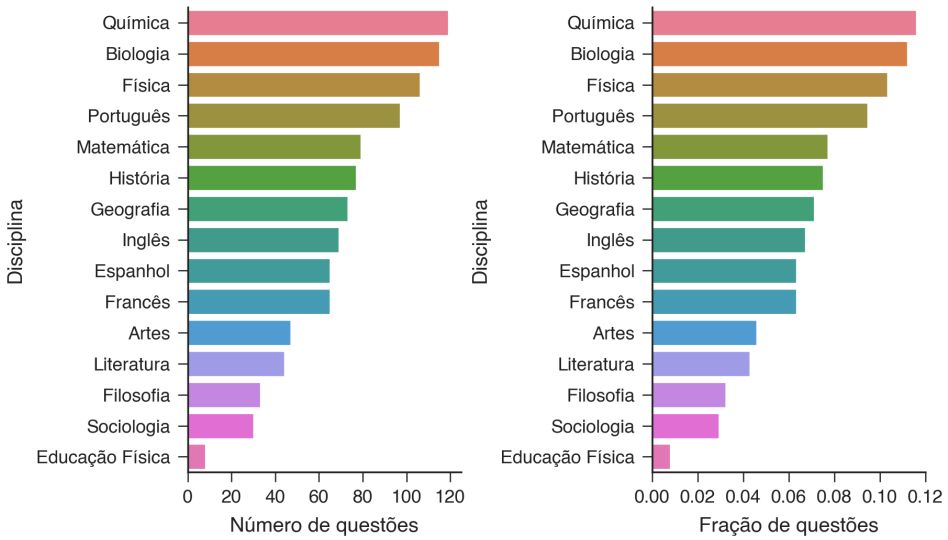


Gráfico 2: Distribuição das questões entre as disciplinas.

Mesmo havendo um formato de interdisciplinaridade nas provas promovidas pela UEM, existe uma prioridade nas matérias previstas nos editais. Para comparar o formato adotado historicamente nas provas, a Figura 6 foi retirada do Edital N.º 031/2015-CVU

7.2 **Composição e avaliação das provas**

7.2.1 A Prova 1 é composta de **40 (quarenta) questões** de alternativas múltiplas, elaboradas na perspectiva interdisciplinar, envolvendo conteúdos referentes às seguintes matérias do Ensino Médio: Arte, Biologia, Filosofia, Física, Geografia, História, Matemática, Química e Sociologia. O conteúdo dessa prova é o mesmo para todos os candidatos aos cursos de graduação ofertados pela UEM.

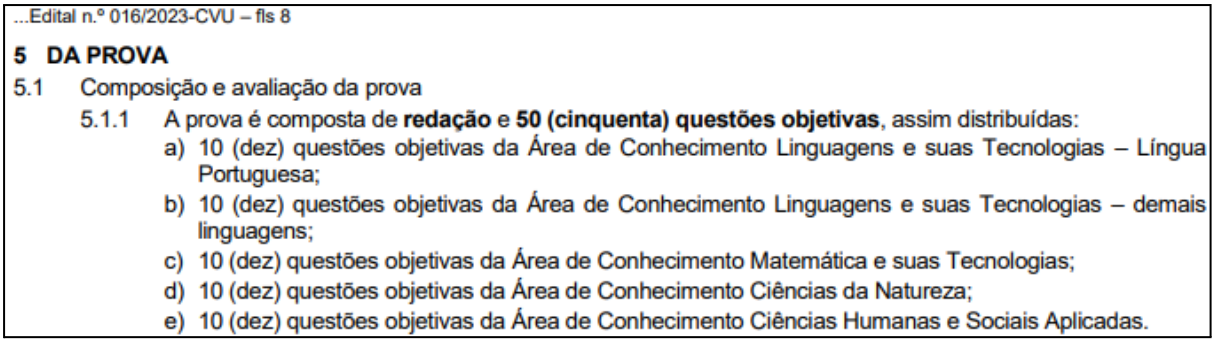
Figura 6. Composição da prova 1 (Conhecimentos gerais) do vestibular de verão da UEM em 2015.

Na época, a prova era dividida em 3 dias, como mostra a Tabela 3.

Dia 1	Prova 1 – Conhecimentos Gerais
Dia 2	Prova 2 – Língua Portuguesa e Literaturas em Língua
Dia 3	Prova 3 – Conhecimentos Específicos

Tabela 3: Divisão de provas no vestibular de verão da UEM em 2015.

Essa divisão permite uma separação mais explícita das linguagem que seria abordada em cada prova. Oito anos depois, no vestibular de verão de 2023 (Figura 7) de Edital N.º 016/2023-CVU, encontramos as informações mostradas na Figura 7.



...Edital n.º 016/2023-CVU – fls 8

5 DA PROVA

5.1 Composição e avaliação da prova

5.1.1 A prova é composta de **redação** e **50 (cinquenta) questões objetivas**, assim distribuídas:

a) 10 (dez) questões objetivas da Área de Conhecimento Linguagens e suas Tecnologias – Língua Portuguesa;
b) 10 (dez) questões objetivas da Área de Conhecimento Linguagens e suas Tecnologias – demais linguagens;
c) 10 (dez) questões objetivas da Área de Conhecimento Matemática e suas Tecnologias;
d) 10 (dez) questões objetivas da Área de Conhecimento Ciências da Natureza;
e) 10 (dez) questões objetivas da Área de Conhecimento Ciências Humanas e Sociais Aplicadas.

Figura 7. Composição da prova do vestibular de verão da UEM em 2015.

A principal mudança identificada no formato do vestibular de verão da UEM em 2023, em relação ao modelo adotado em 2015, foi a reorganização das questões em grandes áreas, mudança que gerou bastante impacto na matemática que agora possui uma divisão própria. Além disso, a prova passou de um formato distribuído em três dias para a realização em apenas um dia, com uma estrutura mais próxima do modelo adotado pelo ENEM. Essa nova organização estabelece a divisão das 50 questões objetivas entre Linguagens e suas Tecnologias, Ciências da Natureza, Ciências Humanas e Matemática, refletindo uma tentativa de modernizar e alinhar o vestibular a critérios de avaliação sem perder a característica da interdisciplinaridade.

A reorganização das questões no vestibular da UEM reflete a evolução das estratégias de avaliação ao longo dos anos. Essa reestruturação não se limita ao formato geral da prova, mas também se reflete em aspectos mais sutis, como a construção das alternativas e o uso de padrões textuais. Como forma de procurar uma diferenciação não explícita entre as alternativas verdadeiras e falsas, o Gráfico 3 representa a distribuição do número de caracteres entre essas alternativas.

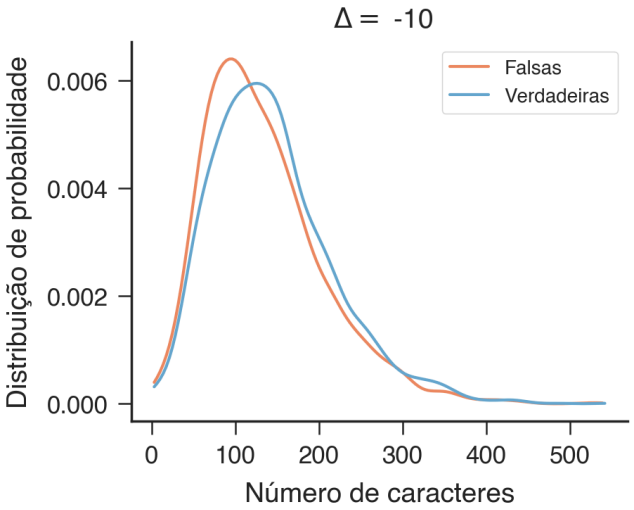
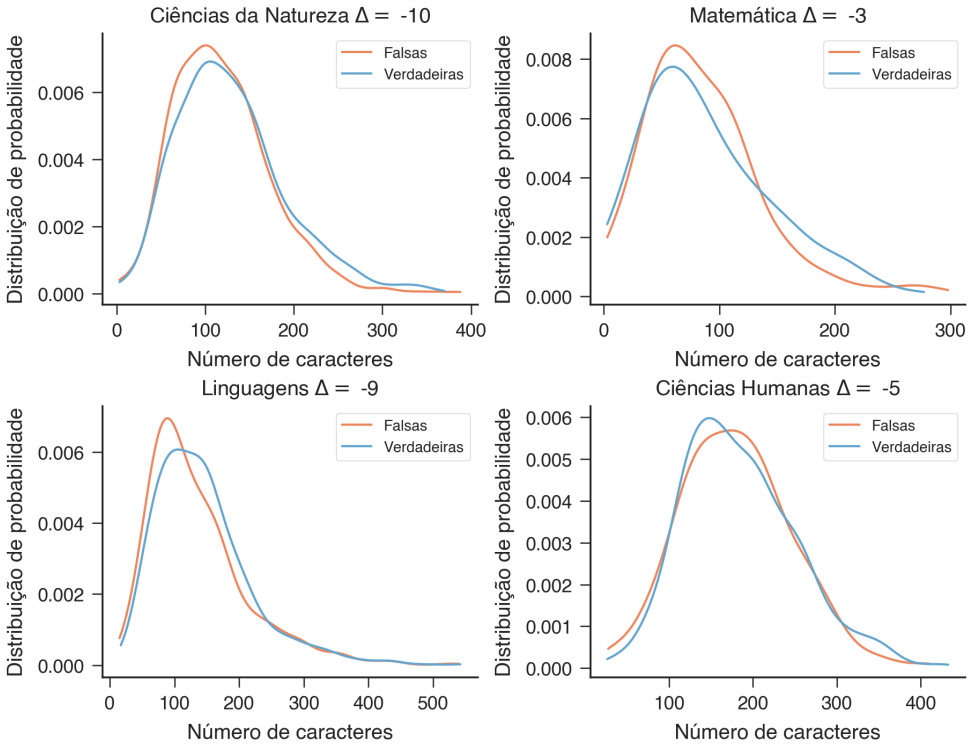


Gráfico 3: Distribuição do número de caracteres das alternativas falsas e verdadeiras em todas as áreas contidas nas provas da UEM.

As alternativas falsas apresentam, em média, Δ=10 caracteres a menos do que as verdadeiras. Embora essa diferença seja sutil e insuficiente para permitir a distinção clara entre alternativas corretas e incorretas apenas pela contagem de caracteres, ela pode refletir um padrão textual que ocorre de forma quase acidental durante a elaboração das questões.



Gráficos 4: Distribuição de probabilidade em comparação com número de caracteres na divisão entre Ciências da Natureza, Humanas, Linguagens e Matemática.

Mesmo quando separamos as questões por grandes áreas (Gráfico 4) ou até mesmo por matérias isoladas (Gráficos 5 e 6) e realizamos novamente a análise em relação ao número de caracteres, não é observada uma variação significativa. O único caso em que ocorre uma maior variação (Δ=28 caracteres) é na matéria de educação física, que começou a ter uma questão exclusiva nas provas somente após o vestibular de 2021. Por ter começado a pouco tempo, a quantidade de dados presente na análise pode ter sido responsável por gerar essa diferença maior.

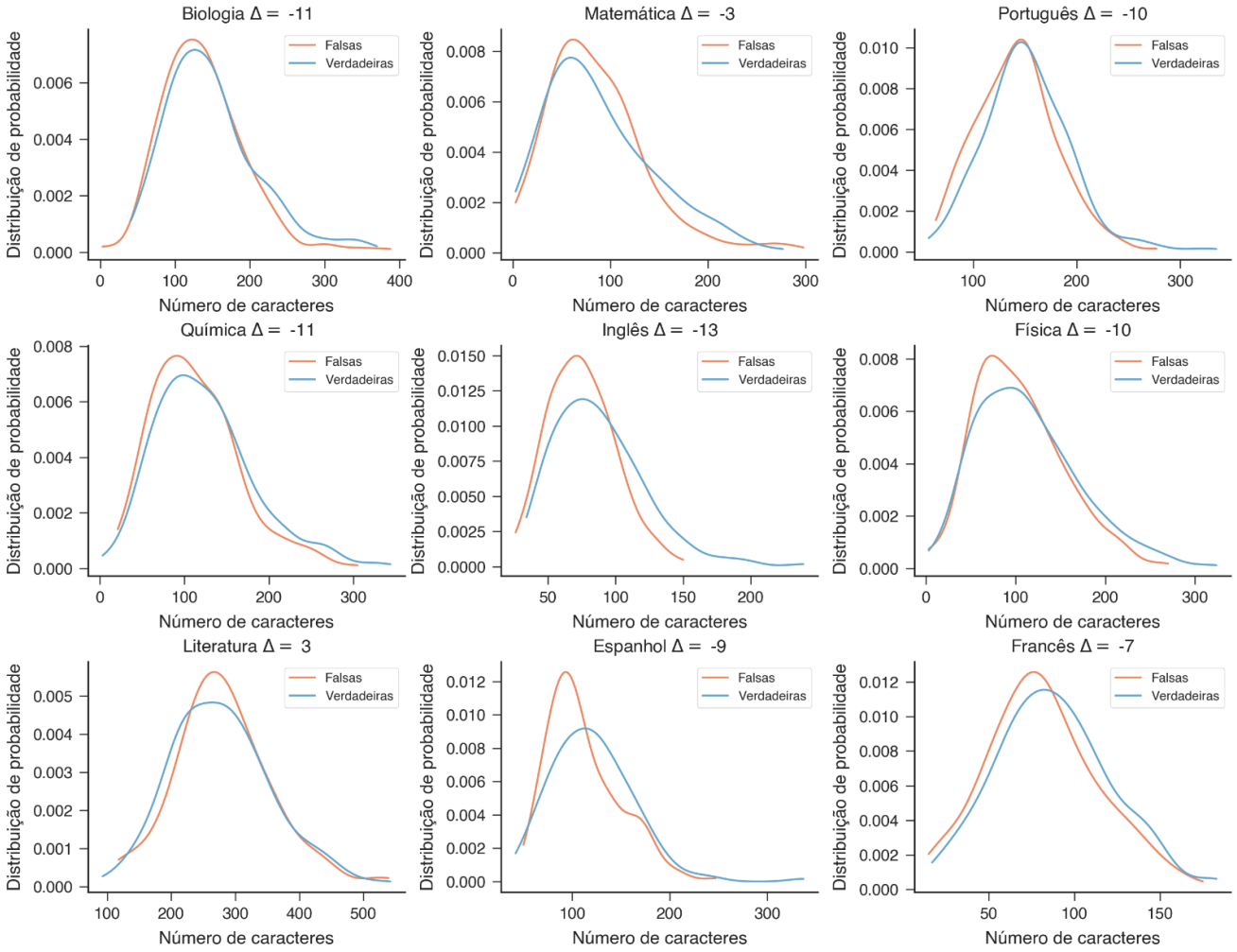


Gráfico 5: Distribuição de probabilidade em comparação com número de caracteres na divisão entre matérias presentes na prova.

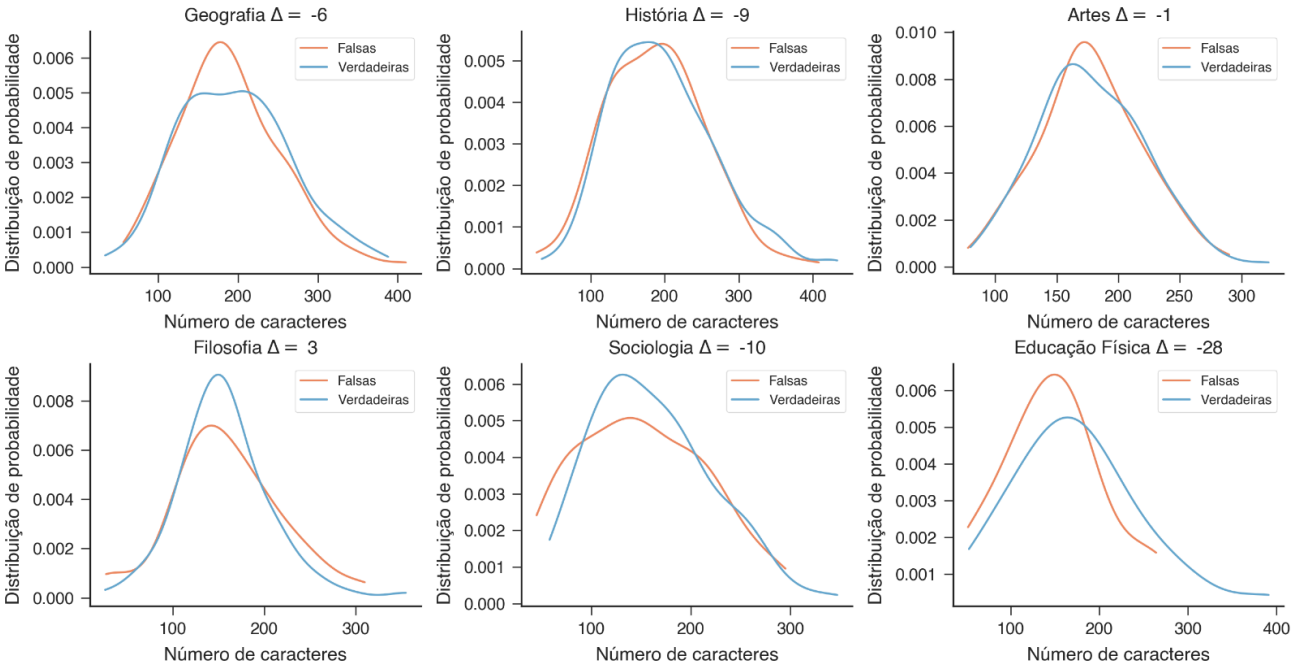


Gráfico 6: Distribuição de probabilidade em comparação com número de caracteres na divisão entre matérias presentes na prova.

Não sendo possível identificar grandes diferenças na quantidade de caracteres nas alternativas, buscamos mudar o foco da distribuição de probabilidade, adotando a análise dos sentimentos dos textos. Em uma das abordagens foi utilizado o VADER (*Valence Aware Dictionary and sEntiment Reasoner*) pois não requer o treinamento de um modelo, podendo ser facilmente utilizado via bibliotecas de código aberto em Python, como o LeIA (Léxico para Inferência Adaptada) para português. O léxico é uma coleção de palavras associadas a valores que indicam suas características emocionais, como positiva, negativa ou neutra. Por exemplo, palavras como "alegria" recebem valores positivos, enquanto "raiva" recebe valores negativos. O princípio de funcionamento é objetivo: no léxico, cada termo possui um valor previamente definido, ao processar os textos dos vestibulares, é gerado um dicionário com uma avaliação de polaridade, calculada a partir dos valores das palavras presentes. Esse dicionário inclui um valor geral de sentimento normalizado (*compound*), que varia de -100 (muito negativo) a +100 (muito positivo) **(22)**. Esse índice permite identificar o sentimento predominante no texto, utilizando os seguintes intervalos como referência:

Sentimento positivo: *compound* >= 5
Sentimento negativo: *compound* <= -5
Sentimento neutro: (*compound* > -5) e (*compound* < 5)

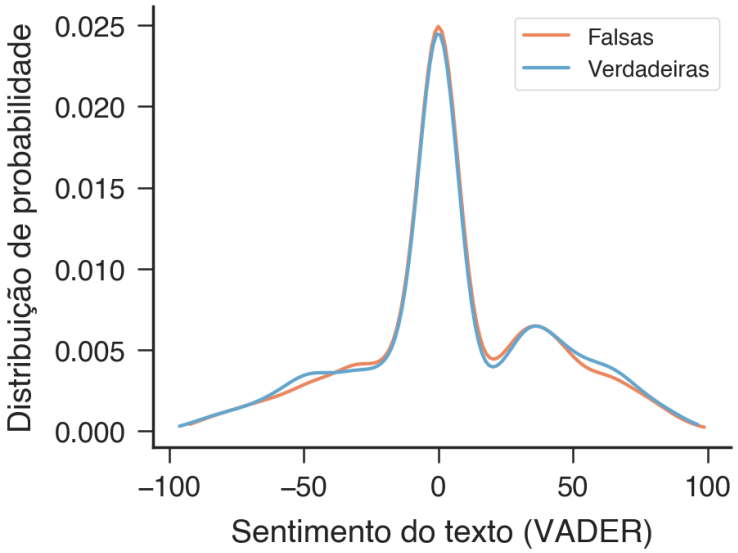


Gráfico 7: Distribuição de probabilidade em relação ao sentimento do texto utilizando o modelo VADER.

Como as curvas apresentaram comportamentos muito semelhantes no Gráfico 7, não foi possível identificar diferenças significativas na probabilidade de as alternativas verdadeiras e falsas se distinguirem em relação aos seus sentimentos tomando como base o VADER. Diante disso, decidiu-se utilizar outro modelo para verificar e confirmar esses resultados (Gráfico 7).

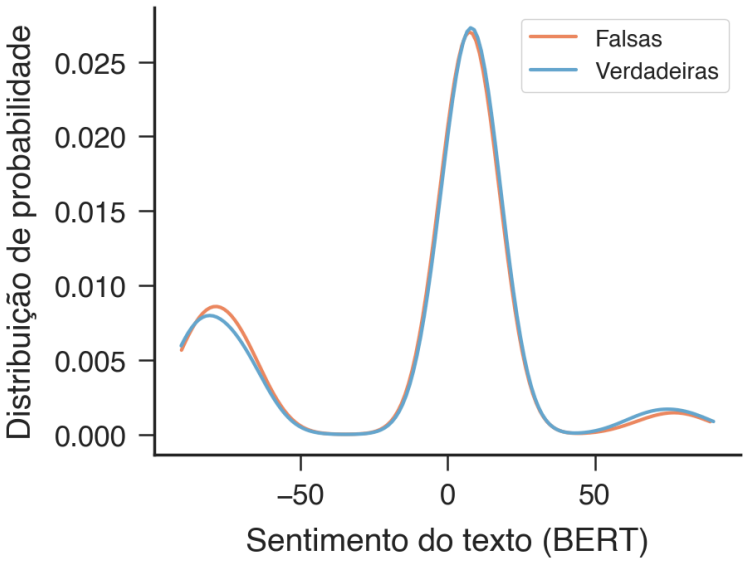


Gráfico 8: Distribuição de probabilidade em relação ao sentimento do texto utilizando o modelo BERT.

Nesta segunda análise de sentimentos, foi adotado o modelo FinBERT-PT-BR, uma ferramenta ajustada ao português brasileiro especializada em textos do contexto financeiro

(23). O modelo foi desenvolvido a partir do fine-tuning do BERTimbau, um modelo BERT geral para a língua portuguesa treinado com uma base de dados de 1,4 milhão de textos relacionados ao mercado financeiro. Embora o FinBERT-PT-BR tenha sido treinado especificamente em dados financeiros, a análise de sentimentos lida com aspectos subjetivos, como emoções e opiniões. Por isso, sua aplicação em contextos fora do domínio financeiro, como os textos de vestibulares, pode ser interessante, mesmo que o modelo não seja o mais indicado para esse tipo de tarefa.

Novamente as curvas do Gráfico 8 não demonstram uma diferenciação na distribuição de probabilidades entre as alternativas verdadeiras e falsas.

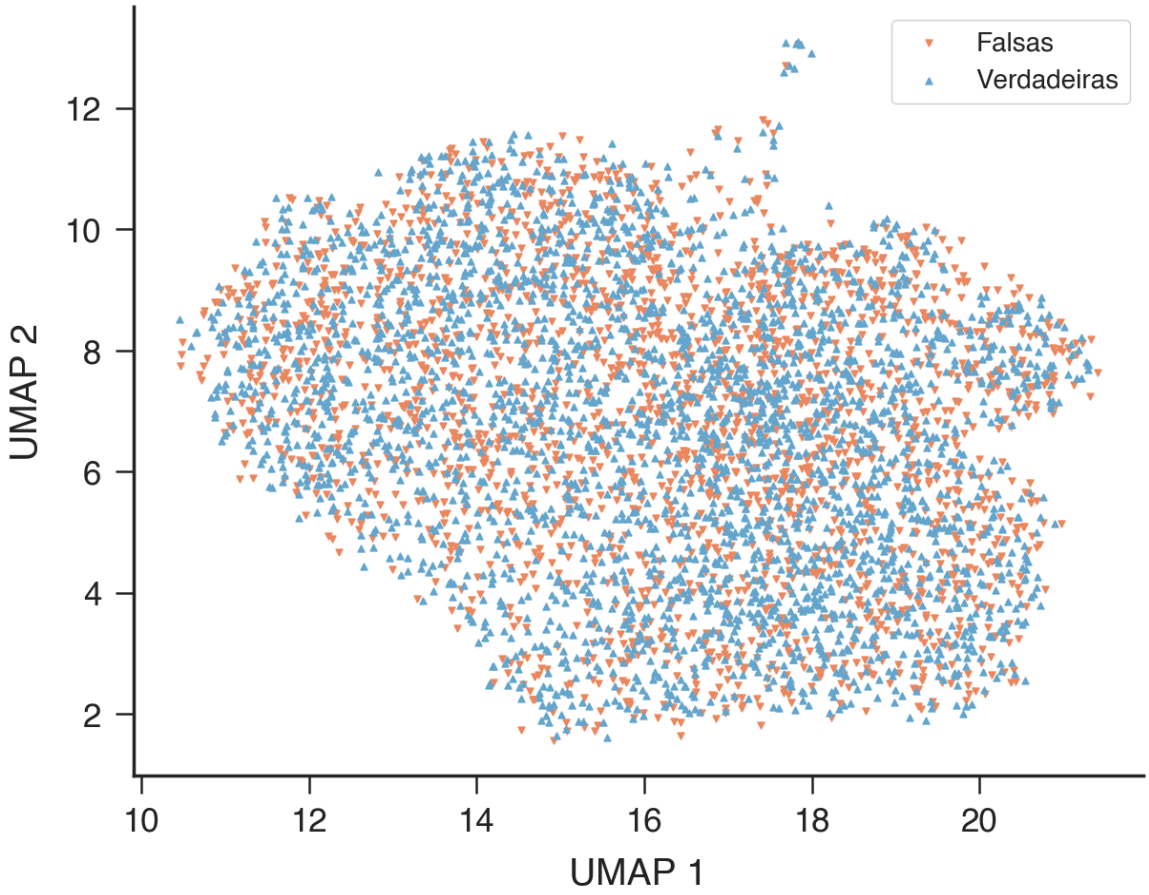


Gráfico 9: Mapa de distribuição vetorial das alternativas verdadeiras e falsas. O gráfico refere-se a uma projeção bidimensional do espaço dos vetores originais por meio do UMAP.

Mapeando as alternativas vetorialmente no espaço utilizando o mesmo modelo BERT (Gráfico 9), observa-se que não há formação de clusters em regiões específicas, com os dados amplamente dispersos e sem agrupamentos visíveis. Essa dispersão sugere a ausência de padrões estruturais que diferenciam alternativas verdadeiras e falsas no nível vetorial

analisado. A distribuição uniforme dos pontos indica que as alternativas, em termos de construção textual, não seguem um padrão sistemático que possibilite sua separação clara.

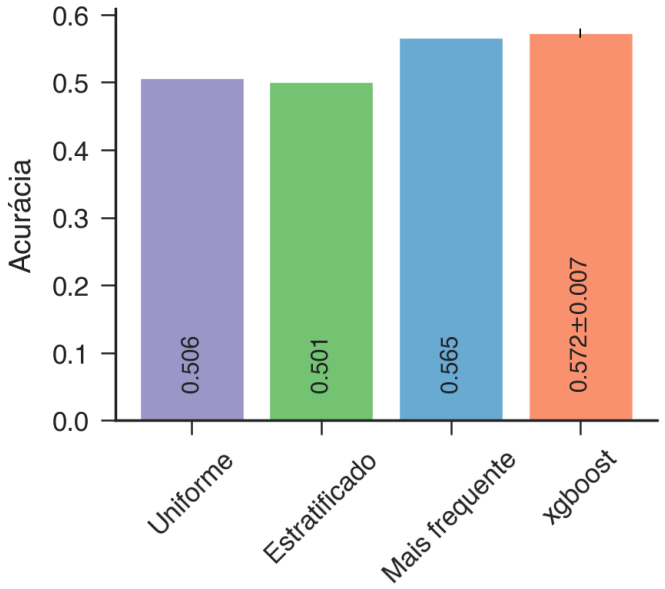


Gráfico 10: Acurácia de predição de acerto quando analisado em panorama geral, dividido entre o modelo XGBoost, e modelos de linha de base.

Ainda que visualmente não seja possível identificar uma diferença, estatisticamente o modelo consegue chegar em uma acurácia de 57,2% (Gráfico 10) de acerto quando submetido a análise de alternativas verdadeiras e falsas representando uma taxa ligeiramente maior quando consideramos linhas de base uniforme, estratificadas ou que sempre assinalam a resposta mais frequente.

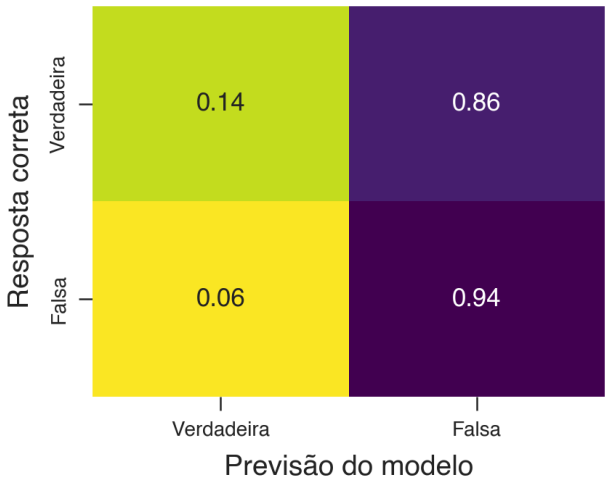


Gráfico 11: Matriz de confusão da porcentagem de acerto e erros da previsão do modelo.

Fazendo a matriz de confusão (Gráfico 11), a taxa de acerto em classificar as alternativas falsas como falsas de fato (também chamado de verdadeiro negativo) é maior do que a taxa de acerto em classificar as alternativas verdadeiras como verdadeiras (também chamado de falso negativo).

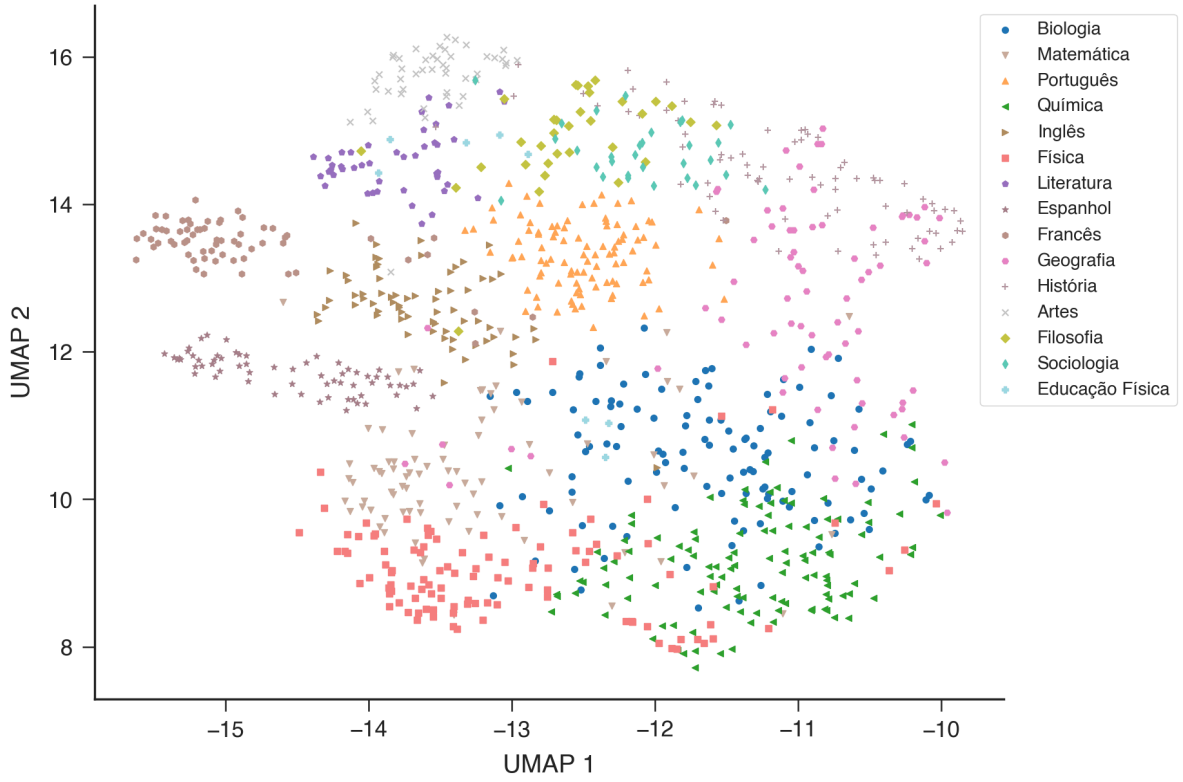


Gráfico 12: Mapa de distribuição vetorial das matérias. O gráfico refere-se a uma projeção bidimensional do espaço dos vetores originais por meio do UMAP.

A análise do mapeamento vetorial das questões por matéria (Gráfico 12), revela regiões específicas no espaço onde cada tipo de questão tende a se concentrar. No entanto, há áreas em que os pontos se sobrepõem ou 'invadem' as regiões de outras matérias, refletindo a interdisciplinaridade da prova. Essa característica é comum em questões cujo contexto pode ser relacionado a mais de uma área do conhecimento. Um exemplo é a proximidade entre as questões de Matemática e Física, bem como entre Biologia e Química, áreas que frequentemente compartilham temas ou abordagens semelhantes. Em contraste, as questões de Francês formam um *cluster* mais isolado, se misturando pouco com outras matérias, podendo ser explicado pelo fato de o Francês ser uma língua estrangeira distinta, sem conexões diretas com o restante do conteúdo abordado no exame.

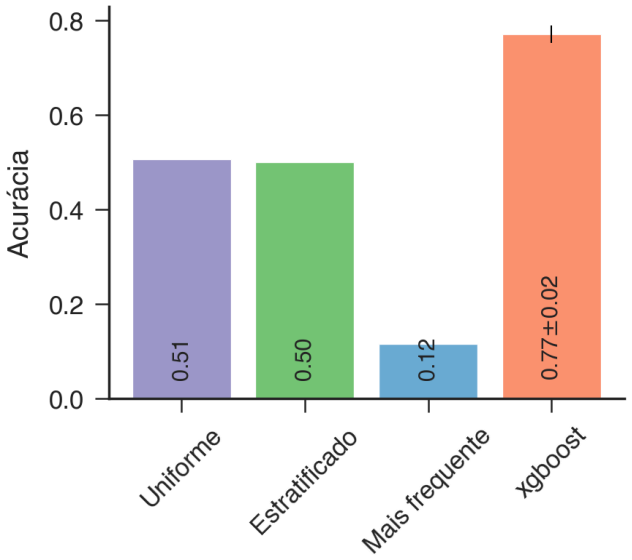


Gráfico 13: Acurácia de predição de acerto quando analisada a capacidade de classificar em tipo de matéria, dividido entre o modelo XGBoost, e modelos de linha de base.

Neste modo de separação por matérias, a acurácia do modelo alcança 77%, um valor muito mais significativo em relação aos modelos de linha de base (Gráfico 13).

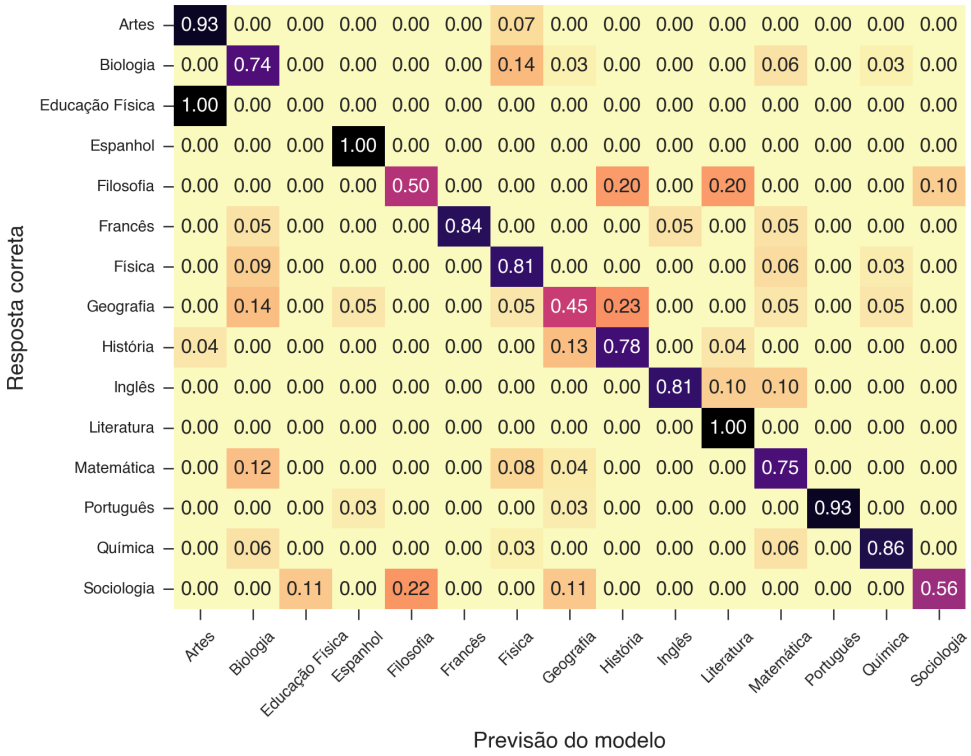


Gráfico 14: Matriz de confusão quando analisada por matérias.

A matriz de confusão entre as matérias (Gráfico 14), proporciona uma compreensão mais clara das relações existentes entre elas. Um dos casos em que ocorreu uma maior confusão, foi a classificação de questões de Filosofia, História, Literatura e Sociologia. Todas, pertencentes às Ciências Humanas, compartilham como foco principal o estudo da linguagem humana e da sociedade. Devido à forte intersecção temática entre essas áreas, é comum que questões de uma disciplina abordem conteúdos que poderiam, facilmente, ser associados a outra.

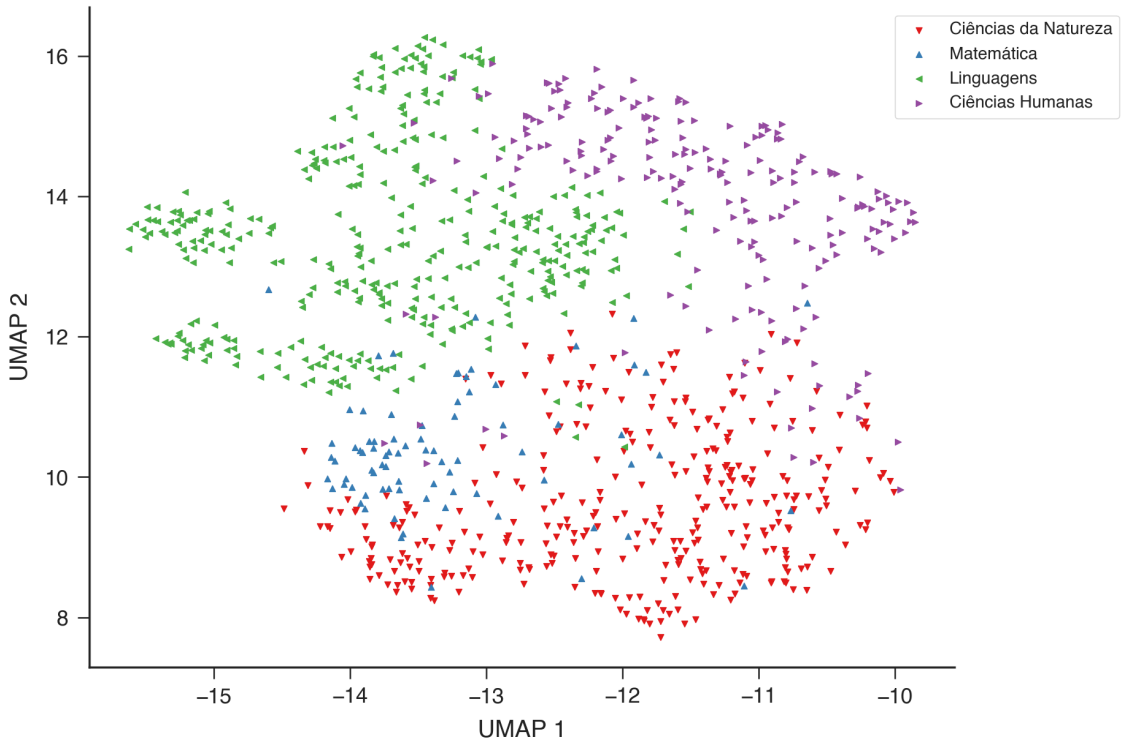


Gráfico 15: Mapa de distribuição vetorial das questões divididas por grandes áreas. O gráfico refere-se a uma projeção bidimensional do espaço dos vetores originais por meio do UMAP.

Assim como no mapeamento de matérias individuais, a separação por grandes áreas se mostra efetiva (Gráfico 15). Ciências da Natureza acabam sofrendo intersecções frequentes com a Matemática, enquanto que Linguagens e Ciências Humanas apresentam regiões mais isoladas.

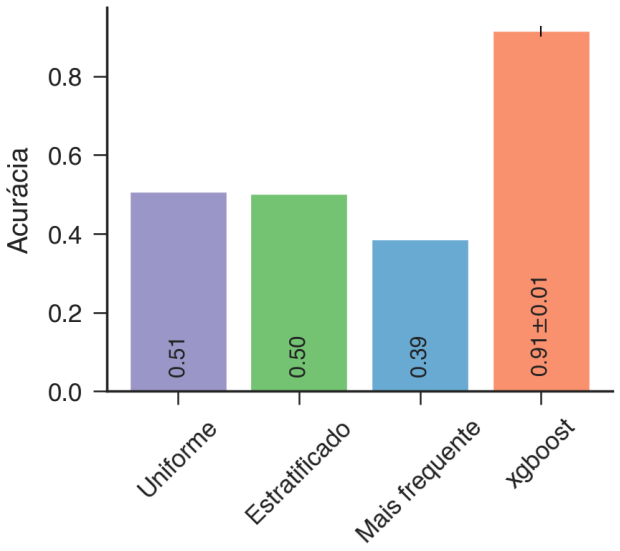


Gráfico 16: Acurácia de predição de acerto quando analisada a capacidade de classificar em grandes áreas, dividido entre o modelo XGBoost, e modelos de linha de base.

Quando a análise é restringida às grandes áreas que compõem as provas do vestibular da UEM, o modelo apresenta um desempenho ainda mais elevado, atingindo uma taxa de acerto de 91% na classificação das questões (Gráfico 16). Em comparação com a taxa esperada para classificações aleatórias ou baseadas na resposta mais frequente, observa-se um ganho significativo de 41% na precisão, o que comprova a eficiência do modelo em distinguir as grandes áreas do exame.

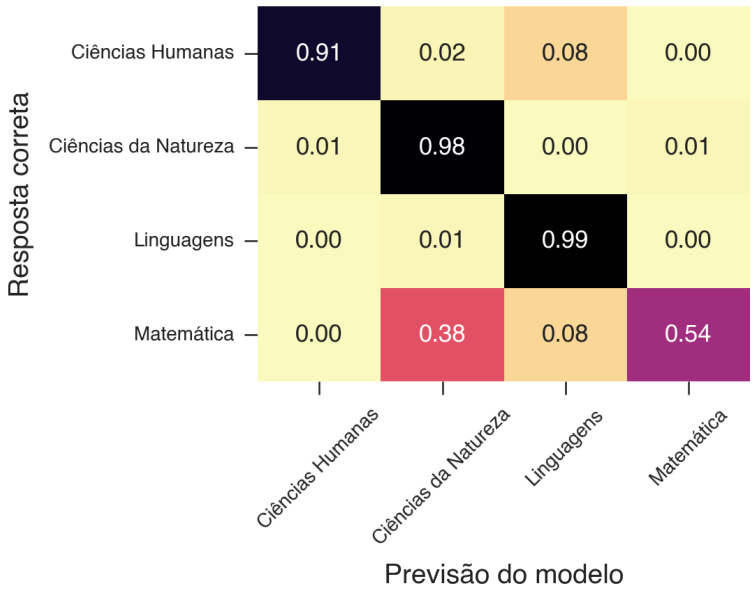


Gráfico 17: Matriz de confusão da porcentagem de acerto e erros da previsão do modelo quando analisado por grandes áreas.

A matriz de confusão (Gráfico 17) aplicada às grandes áreas complementa a análise espacial apresentada no Gráfico 15, traduzindo em números a separação entre as diferentes áreas do conhecimento. A maior parte das questões é corretamente atribuída às suas respectivas áreas, com exceção de Matemática, que se destaca como uma das disciplinas com maior potencial de interdisciplinaridade, possivelmente justificando a maior dificuldade do modelo em classificá-la de forma precisa.

5. CONCLUSÃO

Como esperado, as provas de vestibular da UEM seguem padrões linguísticos naturais e previsíveis, em que as palavras mais frequentes pertencem à categoria das funções gramaticais básicas, como preposições, artigos e conjunções, demonstrando coerência estrutural nos textos e aderência a características universais da linguagem humana. Esses resultados são úteis para análises futuras, especialmente em estudos que visam medir a dificuldade de leitura, a acessibilidade textual e o impacto do vocabulário em avaliações. Entretanto, a tarefa de identificar padrões textuais que diferenciam as alternativas verdadeiras das falsas revelou-se desafiadora. Por exemplo, mesmo quando segmentada por grandes áreas ou disciplinas específicas, a análise do número de caracteres não apresentou variações significativas que indicassem um padrão claro. De maneira similar, a análise de sentimentos, utilizando tanto o modelo VADER quanto o FinBERT-PT-BR, não revelou diferenças marcantes na distribuição de probabilidades entre alternativas corretas e incorretas. Além disso, o mapeamento vetorial das alternativas exibiu ampla dispersão, sem formação de clusters distintos que permitissem separar claramente as alternativas verdadeiras das falsas.

Apesar dessas observações, o modelo BERT demonstrou um desempenho moderado, alcançando uma acurácia de 57,2% na identificação de alternativas corretas e incorretas, ligeiramente superior à probabilidade esperada para um palpite aleatório ou resposta mais frequente. Notavelmente, quando a tarefa foi adaptada para classificar as questões por disciplinas e grandes áreas, a acurácia do modelo atingiu 77% e 91%, respectivamente. Isso ressalta o potencial do modelo para identificar padrões gerais de classificação, especialmente em áreas interdisciplinares, como Matemática e Ciências Humanas, conforme evidenciado pela análise da matriz de confusão.

Com base nos resultados obtidos, conclui-se que as provas de vestibular da UEM não apresentam vieses textuais sistemáticas que favoreçam a identificação das alternativas corretas com base em características linguísticas ou estruturais. Essa ausência de padrões previsíveis enfatiza a importância da aleatoriedade na construção das provas, garantindo a justiça e a confiabilidade do processo seletivo. Por fim, sugere-se que estudos futuros explorem outras técnicas de processamento de linguagem natural e abordagens complementares, como a análise de relações semânticas mais profundas, a inclusão de novos corpora e um treinamento de modelos específicos para lidar com provas avaliativas a fim de expandir nossa compreensão sobre os padrões linguísticos presentes em avaliações. Essas

investigações podem contribuir não apenas para o aprimoramento dos processos, mas também para o desenvolvimento de ferramentas educacionais inovadoras.

6. REFERÊNCIAS

(1) Andrade, L. M. (2000). A escrita, uma evolução para a humanidade. *Linguagem em (Dis) curso*, *1*(1).

(2) Reis, C. K. (2019). *História da escrita: uma contextualização necessária para o processo de alfabetização*.

(3) Briggs, A., & Burke, P. (2016). *Uma história social da mídia: de Gutenberg à internet* (3a ed.). Jorge Zahar Editor.

(4) Mollica, M.C., Quadrio, A.C., Batista, H.R., Maia, M.V., & Leal, M.B. (2018). *Lendo pelo olho mágico.*

(5) OLIVER, É. V. (2007). *Mário, the Brazilian and Manuel, the Lusitanian: Notes on the Brazilian Language in the Correspondence between Mário de Andrade and Manuel Bandeira. Portuguese Studies,* 23 (2), 167–190.

(6) Zipf, G. (1949). *Human Behavior and the Principle of Least Effor*t. New York: Addison-Wesley;

(7) Lima, M.A., Silva, E.A., & da Silva, A.S. (2024). *Um Estudo sobre o uso de Modelos de Linguagem Abertos na Tarefa de Recomendação de Próximo Item*. Anais do XXXIX Simpósio Brasileiro de Banco de Dados (SBBD 2024).

(8) Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

(9) Carvalho, A. P. de L. F. de. *Redes neurais artificiais*. ICMC - USP. Disponível em https://sites.icmc.usp.br/andre/research/neural/

(10) McCulloch, W. S., & Pitts, W. (1990). *A logical calculus of the ideas immanent in nervous activity.* Bulletin of Mathematical Biology, 52(1-2), 99-115.

(11) LIMA, T. S. de, LEMOS, J. F., & LEMOS, J. F. (2022). *Implementação e aplicação do algoritmo backpropagation nas redes neurais artificiais perceptron de múltiplas camadas.* Anais do III CoBICET - Congresso Brasileiro Interdisciplinar em Ciência e Tecnologia, (pp. 1-14)

(12) *Hadi, M. U. et.al (2023). A survey on large language models: Applications, challenges, limitations, and practical usage. TechRxiv. https://doi.org/10.36227/techrxiv.23589741.v1*

(13) AN, J.; DING, W.; LIN, C. *ChatGPT. tackle the growing carbon footprint of generative AI*, v. 615, p. 586, 2023.

(14) GRIGOLETTI, Pablo Souza. *Cadeias de Markov, v. 19, n. 10, p. 2014, 2011.*

(15) SANTANA, D. S. *Assistência bibliográfica durante a escrita de textos científicos: uma abordagem com modelos de linguagem pré-treinados*. 2020.

(16) CHURCH, K. W. (2017). Word2Vec. *Natural Language Engineering*, *23*(1), 155–162. doi:10.1017/S1351324916000334

(17) MIKOLOV, T; CORRADO, G; CHEN, K; DEAN, J. (2013). *Efficient Estimation of Word Representations in Vector Space. Em Proceedings of the International Conference on Learning Representations,* 2013-a.

(18) ALUÍSIO, S. M., & ALMEIDA, G. M. de B. (2021). *O que é e como se constrói um corpus? Lições aprendidas na compilação de vários corpora para pesquisa linguística.* Calidoscópio, 4(3), 156–178.

(19) PETERS et al. (2018). *Deep contextualized word representations. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Vol. 1, pp. 2227-2237). Association for Computational Linguistics. https://doi.org/10.18653/v1/N18-1202.

(20) Devlin et al. (2018). *BERT: Pre-training of deep bidirectional transformers for language understanding*. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423.

(21) Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing.* MIT Press.

(22) Santos, V., Freitag, R.M., Tejada, J., Florêncio, A., Souza, P., & Gois, T.S. (2023). *Sentiment analysis with emojis: a model for Brazilian Portuguese.* International Conference on Language, Data, and Knowledge.

(23) Santos, L., Bianchi, R., & Costa, A. (2023). *FinBERT-PT-BR: Análise de Sentimentos de Textos em Português do Mercado Financeiro*. In Anais do II Brazilian Workshop on Artificial Intelligence in Finance, (pp. 144-155). Porto Alegre: SBC. https://doi.org/10.5753/bwaif.2023.231151.