Experimentos de Análise de Sentimentos em Tweets sobre os cursos de Universidades Públicas

Gabriel Correia de Vasconcelos
Departamento de Ciência da Computação
Universidade de Brasília
Campus Universitário Darcy Ribeiro, DF
gcvasconcelos98@gmail.com

Raphael Luís Souza de Queiroz
Departamento de Ciência da Computação
Universidade de Brasília
Campus Universitário Darcy Ribeiro, DF
rapha95@gmail.com

***Resumo*—Projeto final da disciplina de Fundamentos de Sistemas Inteligentes com objetivo de firmar os conceitos fundamentais de Sistemas Inteligentes, bem como desenvolver habilidades na modelagem computacional de problemas na área, avaliando projetos de pesquisa e conferências e incrementando inovações tecnológicas a linhas de pesquisa existentes. O projeto consiste na modelagem de uma solução computacional, utilizando sistemas inteligentes, para a resolução de um problema identificado pela equipe.**

***Index Terms*—projeto de sistemas inteligentes, processamento de linguagem natural, analise de sentimentos, python**

I. INTRODUÇÃO

A. Descrição do Problema

Um grande problema que universidades públicas vem enfrentando são altos índices de evasão dos alunos. Segundo o INEP (Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira), a taxa de evasão no ensino superior atual gira em torno de 21%, o que corresponde a mais de um milhão de estudantes [1], o que é número preocupante, tanto na perda de mão de obra qualificada para o país, como no prejuízo que isso causa aos cofres públicos. De acordo com um relatório técnico do MEC, em média, um aluno em universidade federal custou, em 2016, R$ 3.129 por mês (cerca de R$ 37.551 no ano) [2].

Como a tabela da figura 1 mostra, existem cursos em que a situação da evasão é ainda mais alarmante, como os cursos de computação, matemática e engenharias.

Existem vários fatores que influenciam a decisão de evasão de um aluno, mas o fato é que ainda muito pouco é feito dentro da universidade para reverter esse quadro, e o que é feito, não está se mostrando eficiente, visto que a porcentagem de 21% se mantém quase constante há mais de 10 anos [1].

Com os recentes anúncios do Ministro da Educação de cortes no sistema de educação público, é essencial que a universidade como instituição e seus departamentos tomem atitudes concretas para usar seus recursos de forma mais eficiente e a diminuir esses altos índices de evasão. Mas para isso são necessários dados confiáveis e relevantes que os ajudem a tomar decisões mais direcionadas.

Tabela IV por área	Evasão 2014/2015				
Agricultura, florestas e recursos pesqueiros	16%	**Direito**	17%	**Proteção Ambiental**	16%
Pública	15%	Pública	5%	Pública	15%
Privada	19%	Privada	18%	Privada	38%
Arquitetura e construção	21%	**Engenharia e Profissões correlatas**	23%	**Saúde**	20%
Pública	7%	Pública	12%	Pública	8%
Privada	24%	Privada	27%	Privada	22%
Artes	15%	**Formação de Professor e Ciências da Educação**	19%	**Serviço Social**	19%
Pública	4%	Pública	16%	Pública	11%
Privada	22%	Privada	23%	Privada	21%
Ciências	25%	**Humanidades e Letras**	19%	**Serviços de Segurança**	6%
Pública	10%	Pública	17%	Pública	6%
Privada	31%	Privada	23%	Privada	10%
Ciências Físicas	18%	**Jornalismo e Informativo**	28%	**Serviços de Transporte**	25%
Pública	17%	Pública	18%	Pública	-
Privada	24%	Privada	34%	Privada	25%
Ciências Sociais e Comportamentais	20%	**Matemática e Estatística**	30%	**Serviços Pessoais**	26%
Pública	15%	Pública	30%	Pública	17%
Privada	22%	Privada	39%	Privada	33%
Computação	28%	**Produção e Processamento**	23%	**Veterinária**	17%
Pública	22%	Pública	18%	Pública	6%
Privada	31%	Privada	30%	Privada	21%
Comércio e Administração	23%				
Pública	13%				
Privada	24%				

Figura 1: Tabela construída a partir de um estudo do Instituto Lobo para o Desenvolvimento da Educação, da Ciência e da Tecnologia [3]

B. Propósito do Projeto

O propósito maior da equipe é empoderar a universidade e seus departamentos com informações valiosas sobre o comportamento de seus alunos e provar como suas manifestações e redes sociais podem ser gatilhos para a formulação de programas que de fato diminuam a evasão escolar.

Este projeto resolve uma parte inicial e tem o objetivo de validar a hipótese de que a polaridade dos comentários sobre um curso nas redes sociais está ligada aos seus índices de evasão de alunos.

Com o objetivo de segmentar o escopo do projeto para inicialmente validar a ideia e depois permitir a escalabilidade, foi

escolhida como universidade alvo a Universidade de Brasília, que de acordo com um estudo interno recente realizado pelo Decanato de Ensino de Graduação, aponta que quatro em cada 10 estudantes abandonam a instituição [4].

Outra segmentação ocorreu na escolha de grandes áreas de estudo no lugar de se analisar cada curso individualmente. Ou seja, alguns cursos foram agrupados devido a limitações da API utilizada para se montar a base de dados e por motivos de alta complexidade do algoritmo na escolha da opção anterior.

Com esta hipótese validada se torna possível escalar o projeto para todas as universidades e se analisar com mais detalhes cada curso oferecido pela universidade.

C. Referências teóricas

Como base teórica para guiar o desenvolvimento do projeto foram escolhidos dois artigos.

O primeiro encara o desafio de se fazer um estudo comparativo entre diversos algoritmos de análise de sentimentos na língua portuguesa [5], área que ainda está nos estágios iniciais quando comparada a outras línguas, como a inglesa. Neste nos inspiramos nas etapas de pré-processamento dos dados e em como foi analisado o sentimento do texto.

O segundo aplica a análise de sentimentos para tweets de um assunto específico, algo similar ao que nossa projeto está fazendo, mas que no caso deles se refere aos tweets que falam sobre os jogos paraolímpicos de 2016 [6].

Ambos as referencias foram publicadas no ano de 2018.

II. MODELO

A. Ciclo de Vida Analítico

Como metodologia para a abordagem do problema foi escolhida foi escolhido o Ciclo de Vida Analítico desenvolvido pela empresa de inteligência de negócio SAS, a maior companhia de software do mundo e que trabalha com as maiores do mercado em soluções para análise de dados.

Como ilustrado na figura 2, o ciclo se inicia na etapa de Exploração de Dados, e significa entender os dados disponíveis, como estão organizados e o que é possível se extrair deles. Essa fase, para o problema do grupo, consistiu na identificação do Twitter como uma rede social que poderia ser a fonte de nossa base de dados, a exploração dos limites da API que a empresa disponibilizava e a limpeza de uma grande massa de dados para filtrar apenas o que era necessário para a validação da hipótese inicial. Ou seja, a massa de dados foi construída e preparada para a próxima etapa.

Na fase de Desenvolvimento do Modelo acontece a parte analítica do ciclo, ou seja, a criação de modelos preditivos e de classificação, e seleção de dados para teste e treino. No nosso processo, aqui desenvolvemos os algoritmos para analisar os sentimentos dos textos presentes em cada tweet e segmentamos as grandes áreas que seriam analisadas testando diferentes variáveis.

Por último, acontece a etapa de Implantação de Modelo, onde acontece a comparação entre os modelos desenvolvidos e a atribuição de pontuações destes modelos. Como será definido com mais detalhes na seção de solução, foram desenvolvidos

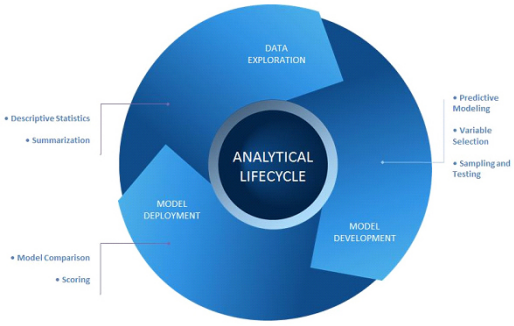

Figura 2: Imagem que ilustra o ciclo de vida analítico

Esta é uma sentença.

['esta', 'é', 'uma', sentença', '.']

Figura 3: Tokenização de uma sentença.

três modelos que foram comparados e somados para melhorar a acurácia do algoritmo final.

A ideia do processo ser cíclico vem justamente da iteração entre as etapas e da continuidade que é necessária para se desenvolver uma boa solução analítica.

A adoção desta metodologia foi essencial para dar um direcionamento no desenvolvimento e um importante ponto de inovação no projeto. Dando um ponto de vista de produto a solução analisada conseguimos nos diferenciar do que normalmente acontece na maioria do que é desenvolvido nos artigos científicos consultados pelo grupo.

B. Processamento de Linguagem Natural

O Processamento de Linguagem Natural (PLN) é uma área da Inteligência Artificial (IA) responsável pelo estudo e entendimento da linguagem do ser humano e fornecer à máquina a capacidade de entender e compor textos. A capacidade de entender um texto significa que a máquina irá poder reconhecer o contexto, extrair informações, analisar sentimentos, entre outros.

O PLN é responsável pelo tratamento computacional de uma ou mais linguagem natural. Para isso, é necessário métodos de pré-processamentos que abstraem a linguagem, deixando apenas as informações que são mais relevantes. Dentre várias, podemos destacar as 2 mais utilizadas neste trabalho:

1) Normalização: A normalização abrange tratativas como a tokenização, transformação de maiúsculas e minúsculas, remoção de caracteres especiais e links, etc. A tokenização tem como função a separação de frases em unidades:

2) Correção Ortográfica: O pré-processamento de Correção Ortográfica é utilizado para tratar um *dataset* que possui erros de digitação, abreviação e vocabulário informal. Esses erros, muito encontrado em textos de redes sociais por

serem informais, podem ser prejudiciais para a análise de um texto, pois geram novos $tokens$ desnecessários, aumentando assim a esparsidade de dados.

C. Análise de sentimentos

A análise de sentimentos, também chamado de opinion mining, tem se tornado comum entre empresas e projetos acadêmicos e tem como principal objetivo a captura de opiniões sobre os mais diversos assuntos, se tornando uma das principais formas de pesquisa atualmente.

A ideia da análise de sentimentos consiste em identificar a opinião expressada em uma determinada frase como negativa, neutra ou positiva. As principais ferramentas utilizadas em análise de sentimentos são as redes sociais, tal como o Twitter que foi utilizado neste trabalho, blogs e mídias sociais. Isto se dá porque essas ferramentas proporcionam um volume imenso de dados acerca de um determinado assunto.

A utilização de algoritmos para análise de sentimentos vem sido utilizado há bastante tempo, sendo iniciado praticamente nos anos 2000 e crescendo desde então. Porém, a quantidade de trabalho nessa área é bem limitada, principalmente devido à complexidade de se estudar um texto vindo de redes sociais, que é um texto bastante informal e desestruturado, fazendo com que o processamento total do texto tenha um custo de tempo muito alto.

III. SOLUÇÃO E ANÁLISE

Para a implementação do algoritmo a equipe escolheu a linguagem *Python*, na versão 3.7.1, devido a sua conhecida facilidade para manipular dados e sua ampla disponibilidade de bibliotecas para nos auxiliar na utilização do algoritmo. A biblioteca principal utilizada foi a *tweepy*, utilizada como interface de acesso a Twitter API e a *json* para a conversão das respostas da API para dicionários, tipo de dado do Python, e possibilitar sua manipulação.

A construção do dataset foi feita a partir da coleta de dados da API (interface de programação de aplicativo) do Twitter onde cada observação é um tweet, uma publicação de algum usuário nessa rede social.

Para se ter acesso a Twitter API é necessário um cadastro a plataforma de desenvolvedores do sistema, onde a empresa oferece um plano gratuito de consumo a API, o plano "Standart", que é o que foi utilizado pela equipe. Este plano possui algumas limitações, pois pode apenas buscar tweets que foram postados em no máximo sete dias e não permite a consulta em lotes grandes, o que dificultou a construção de uma base de dados com mais significância estatística e limitou o escopo do projeto a analisar apenas o contexto atual.

Toda consulta a API utiliza a autenticação "OAuth"e necessita das credenciais fornecidas no cadastro na plataforma previamente citada. Em nossa implementação, estas credenciais estão em um arquivo separado chamado config.py e não estão disponíveis no repositório público devido aos termos de compromisso da plataforma.

A base de dados e constituída por um base de tweets por grandes áreas, cursos específicos e da unb como um todo por meio de palavras-chave presentes nos tweets. Como, após muitas chamadas, a API retorna o erro HTTP 429 ("Too many requests"), o tamanho do dataset foi limitado a por volta de 1500 tweets por conjunto de palavras-chave.

Antes serem inseridos no dataset, são retirados o texto inteiro de cada tweet e seu id, que o identifica unicamente. Esse texto, com até 280 caracteres passa por um processo de limpeza, para que dados irrelevantes a análise de sentimentos não a prejudiquem. São retirados:

- retweets, para evitar repetição de frases no dataset
- arrobas, forma como usuários citam os perfis uns dos outros na plataforma e que são irrelevantes para a análise
- links, novas linhas e pontuações, que prejudicam a identificação de palavras
- internetês, ou seja palavras como 'pq' e 'vc' são trocadas por 'por que' e 'você'

Antes de salvar todos os tweets limpos no dataset, sua polaridade é calculada. Dado alguns problemas que tivemos para a análise de sentimento de cada tweet, serão descritas as abordagens que foram utilizadas a cada tentativa de melhorar o algoritmo.

A. Primeira abordagem

Como primeira abordagem da análise, foi utilizado a função Translator da biblioteca googletrans para o último processamento do tweet e calculado a polaridade deste por meio da função polarity da biblioteca TextBlob.

A biblioteca googletrans foi utilizado para fazer a tradução dos tweets da língua português brasileiro para o inglês antes de calcular a polarização. Este método foi usado para aproveitar a função polarity do TextBlob. Esta função avalia a frase por completo e retorna um valor entre -1.0 e 1.0 a partir da quantidade de termos negativos e positivos presente na frase. Porém, para que esta função seja usada, a frase que serve de input deve estar na língua inglesa.

Esta primeira abordagem se tornou problemática por causa da biblioteca googletrans. Para que o tweet fosse traduzido, a biblioteca manda uma requisição para o Google Tradutor com as línguas escolhidas para tradução e retorna a resposta. Como é necessário fazer a tradução de muitos tweets, a biblioteca envia muitas requisições e o Google bloqueia os IPs por motivos de segurança. Para que a biblioteca seja utilizada, foi necessário utilizar uma VPN quando o IP fosse bloqueado, porém o problema se encontraria mais na frente e iríamos ficar em um cíclo infinito de troca de VPN e bloqueio de IP.

B. Segunda abordagem

Como solução para a primeira abordagem, o grupo decidiu, após a leitura de um outro artigo sobre sentimento de tweets, realizar a análise das palavras individualmente.

Para realizar a segunda abordagem, foi necessário a utilização de um dicionário de palavras na língua portuguesa. Como não havia uma biblioteca em Python ou um dicionário já pronto disponível para ser utilizado no código, o grupo precisou criar o seu próprio. Para isso, foi gerado um arquivo .txt com adjetivos negativos e positivos para o código e

utilizado estes arquivos para a análise individual das palavras encontradas no tweet.

O problema encontrado nesta abordagem se dá pelo motivo de que palavras na língua portuguesa ao serem analisados individualmente podem não possui o mesmo sentido que a palavra dentro de um contexto. Sendo assim, a análise pode estar errado, pois podemos analisar uma palavra que dentro do contexto significa o contrário, gerando assim um falso positivo.

Outro problema encontrado é que um tweet ao ser escrito não reflete 100% do sentimento expressado pela pessoa. Sentimentos como sarcasmo e ironia podem gerar uma análise falsa, pois a análise individual das palavras não auxiliam na detecção desses sentimentos.

C. Terceira abordagem

Para solucionar o problema apresentado na segunda abordagem, foi necessário utilizar a terceira abordagem de analisar os emojis e emoticons presentes nos tweets.

Ao capturar muitos tweets, foi identificado que a grande maioria utiliza emojis para expressar as emoções que os usuários desejam passar ao escrever. Sendo assim, essa abordagem tem como funcionalidade complementar a análise das palavras individuais alinhando essa análise com a análise de emojis e emoticons presentes no mesmo tweet.

A ideia principal da terceira abordagem é continuar analisando o tweet como descrito na segunda abordagem e adicionando a análise de emojis para fazer a correção dá análise feita. Com isso, grande parte dos sarcasmos e ironias poderiam ser identificados e aqueles tweets que geraram falsas análises poderiam ser corrigidos, aumentando assim a eficiência do analisador de sentimentos em poucos porcentos.

IV. RESULTADOS

A análise retorna o valor '-1' se o texto do tweet foi identificado pelo algoritmo com polaridade negativa, '0' se o texto tem polaridade neutra e '1' se tem polaridade positiva.

Na análise dos datasets, somou-se todos os valores de polaridade, de forma valores positivos significavam uma maioria de polaridade geral positiva e valores negativos significavam polaridade geral negativa. Foi formulada uma escala de polaridade para melhor representar a 'intensidade' de polaridade de cada grande área:

- < -15, significa um sentimento geral negativo sobre o objeto analisado
- $[-15, -6]$, significa um sentimento geral ligeiramente negativo sobre o objeto analisado
- $[-5, 5]$, significa um sentimento geral neutro sobre o objeto analisado
- $[6, 15]$, significa um sentimento geral ligeiramente positivo sobre o objeto analisado
- > 15, significa um sentimento geral positivo sobre o objeto analisado

O dataset inicial, construído junto ao desenvolvimento do algoritmo e das abordagens, possui tweets relacionados a UnB, onde foram utilizadas as palavras-chave "unb,Universidade de Brasília,UnB,UNB"na chamada da API. Foram analisados os sentimentos de 3692 tweets e detectado um sentimento geral ligeiramente positivo, resultado dentro das expectativas do grupo. Ao se ler um subconjunto do dataset de 50 tweets, a escolha das palavras-chave se mostrou correta.

O segundo dataset corresponde ao dos cursos de engenharia. Foi utilizada o conjunto de palavras-chave "engenharia,curso de engenharia,FT". Foram analisados 1423 tweets e foi detectado um sentimento geral negativo, resultado também correspondente as expectativas do grupo e alinhado a hipótese apresentada. O uso das palavras não se mostrou tão eficiente devido ao baixo número de tweets em que as pessoas citam seus cursos e expressam um sentimento, bem como o problema das palavras-chave aparecerem em outros contextos, que não o de expressão de um sentimento.

O terceiro dataset se refere ao curso de ciência da computação, um dos identificados com altos índices de evasão. Foram utilizadas as palavras-chave "cic,comp,Ciência da Computação". Foram analisados 1739 tweets e detectado um sentimento geral ligeiramente negativo. O escolha das palavras-chave foi considerada suficientes, principalmente pela escolha dos acrônimos de como os cursos são chamados pelos alunos.

O quarto dataset se refere aos cursos de saúde e utilizou as palavras-chave "medicina,enfermagem,veterinária". Foram 1881 tweets analisados e detectado um sentimento ligeiramente positivo. A escolha das palavras-chave se mostrou correta e o resultado faz sentido, pois estes cursos são os que tem menos evasão, como mostra a figura 1.

O quinto e último dataset se refere a grande área de humanas e utilizou as palavras-chave "filosofia,sociologia,ciência política". Foram 1921 tweets analisados e o sistema detectou um sentimento ligeiramente negativo. Após a análise de um subconjunto do dataset, pôde-se concluir que muitos tweets não são de universitários e que existe um viés muito forte da atualidade, visto a recente discussão sobre a utilidade e eficiência de alguns cursos de humanas.

Foi identificado que alguns dos tweets não pertenciam necessariamente a UnB. Para corrigir isso seria necessário uma segmentação das áreas onde cada tweet foi postado e limitar essa área a região de Brasília e entorno melhoraria os dados captados. Apesar da API possuir esta funcionalidade, ela não foi implementada devido ao seu funcionamento ser muito diferente do que já estava sendo feito no algoritmo e implicaria numa refatoração muito grande.

V. CONCLUSÃO

O problema proposto pelo grupo foi de demonstrar que o sentimento expresso em redes sociais sobre a universidade e seus departamentos estão diretamente ligados com os índices de evasão dos alunos em seus respectivos departamentos. Para que isso fosse demonstrado, o grupo utilizou da rede social Twitter para capturar os tweets postados por alunos e gerar um output que prove este problema.

Verificando os resultados obtidos na sessão acima, podemos verificar que o analisador desenvolvido possui uma acurácia mediana. Isto se dá pelo motivo de que uma máquina não

consegue entender 100% o sentimento que um humano deseja expressar apenas por sua escrita e pelo motivo de que o grupo precisou criar quase que do zero um analisador da língua portuguesa baseado em teorias propostas por artigos científicos, onde a maior dificuldade foi a criação de um dicionário para alimentar a máquina.

Mesmo com um resultado mediano, podemos verificar que o analisador é funcional comparando os resultados obtidos pelo analisador com a tabela apresentada na Figura 1, onde os cursos com maior evasão da Universidade de Brasília são aqueles que apresentaram resultados de sentimentos negativos ou ligeiramente negativos no analisador. Sendo assim, o objetivo do projeto foi alcançado.

Para o futuro, é possível melhorar a acurácia do analisador treinando melhor a máquina com mais dados e melhorando o dicionário disponível para a alimentação da máquina.

REFERÊNCIAS

[1] DA EDUCAÇÃO SUPERIOR, INEP Censo. Notas estatísticas. 2015.

[2] https://oglobo.globo.com/sociedade/entenda-quanto-custa-um-aluno-numa-universidade-federal-brasileira-23666877

[3] https://educacao.estadao.com.br/blogs/roberto-lobo/497-2/

[4] https://www.metropoles.com/distrito-federal/educacao-df/a-cada-10-alunos-que-entram-na-unb-quatro-abandonam-o-curso

[5] DE AGUIAR, Erikson Júlio et al. Análise de sentimento em redes sociais para a língua portuguesa utilizando algoritmos de classificação. In: Anais do XXXVI Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos. SBC, 2018.

[6] DOS SANTOS, Silvan Menezes et al. “Twittando” sobre os Jogos Paralímpicos Rio/2016: Uma análise do Sentimento Paralímpico sob o ponto de Vista de Internautas. 2018.