Análise de sentimentos no *Twitter* para a identificação de indivíduos com sintomas de ansiedade relacionados aos Institutos Federais

Nayane França Gomes[1], Carlos Alexandre Silva[1] , Cristiane Norbiato Targa[1]

[1]Departamento de Informática
Instituto Federal de Minas Gerais (IFMG) – Sabará, MG – Brasil

nayane.ifmg@gmail.com,carlos.silva@ifmg.edu.br, cristiane.targa@ifmg.edu.br

Abstract. *Social networks are increasingly used, either as a form of entertainment, communication or freedom of expression, being responsible for generating large amounts of data, which extracted may present useful knowledge. Several works have been carried out using techniques of knowledge extraction through methods based on machine learning, such as the sentiment analysis, about texts from social networks, allowing to understand several complex social phenomena. One of the social problems that affects thousands of people around the world is anxiety. The objective of this work is the creation of a dictionary for a possible identification of individuals related to the Federal Institutes of Education, Science and Technology of Brazil with the indicators of anxiety. For the identification of these profiles, was developed a dictionary labeled as anxious, using machine learning methods. Tweets of users from 11 Federal Institutes were analyzed according to the model generated.*

Resumo. *As redes sociais são cada vez mais usadas, seja como forma de entretenimento, comunicação ou liberdade de expressão, sendo responsável por gerar grande quantidade de dados, que extraídos podem apresentar conhecimento útil. Diversos trabalhos têm sido realizados utilizando técnicas de extração de conhecimento por meio de métodos baseados em aprendizado de máquina, como a análise de sentimentos, sobre textos oriundos de redes sociais, permitindo compreender diversos fenômenos sociais complexos. Um dos problemas sociais que afeta milhares de pessoas em todo o mundo é a ansiedade. O objetivo deste trabalho é a criação de uma API capaz de gerar e analisar sentenças ansiosas. A ferramenta será usada na identificação de indivíduos relacionados aos Institutos Federais de Educação, Ciência e Tecnologia do Brasil com o indicadores de ansiedade. Para a identificação desses perfis foi desenvolvido um modelo de aprendizado de máquina e um dicionário ansioso contendo 4.975 palavras. Tweets de usuários oriundos de 11 Institutos Federais foram analisados de acordo com o modelo gerado.*

1. Introdução

Alguns autores definem a ansiedade como um estado emocional desagradável, que acarreta em desconfortos somáticos, com relação direta a outra emoção, o medo. Esse estado emocional é geralmente relacionado a um evento futuro e, pode gerar desconfortos como “frio na barriga”, “mãos suadas”, “coração apertado”, e por vezes é descrito como “paralisante” [1].

Pesquisadores têm identificado uma série de cargas crônicas que contribuem para a ansiedade. Dentre elas são citadas: “ser pobre[1], ter histórico de minoria racial, estar desempregado, viver em um lar conturbado e ser solteiro.” Em geral, os transtornos emocionais de estudantes universitários estão relacionados a problemas acadêmicos e sociais como sentimento de solidão, falta de sentido, e necessidade de corresponder a altos padrões [2].

A ansiedade pode se apresentar de formas distintas como, ansiedade enquanto estado ou ansiedade enquanto traço. A ansiedade-estado, se refere a um estado emocional transitório, composto por sentimentos subjetivos de tensão que podem variar em intensidade ao longo do tempo. A ansiedade-traço, se refere a uma disposição pessoal, relativamente estável, a responder com ansiedade a situações estressantes e uma tendência a perceber um maior número de situações como ameaçadoras [3].

Algumas características relevantes podem ser observadas em participantes de pesquisas sob estado de alta ansiedade, onde exibem uma série de problemas relacionados com a aprendizagem [2]:

- Dificuldade de prestar atenção e alto nível de distração. Quando deveriam estar prestando atenção, estão concentradas em sentimentos de inadequação ou pânico, desempenho dos outros, dor de cabeça, de estômago, e fracasso.
- Ficam menos cientes das implicações e complexidades sendo mais propensos a interpretar erroneamente aquilo que leem, especialmente se o material é difícil ou ambíguo.
- Não organizam e elaboram apropriadamente as informações processadas, como fazem as pessoas menos ansiosas.
- Menos adaptáveis quando o processo de aprendizagem exige flexibilidade.

As especificidades ainda não são compreendidas por pesquisadores, mas sabe-se que indivíduos ansiosos possuem dificuldade em acessar as informações na memória. Estudantes que se consideram ansiosos, declaram bloqueio em exames, relatando que mesmo com o conteúdo em mente, o acesso à ele não é tão trivial [2].

---

[1] GASPARIN, Gabriela. **Veja diferenças entre definições de classes sociais no Brasil**. São Paulo, 2018. Disponível em: http://g1.globo.com/economia/seu-dinheiro/noticia/2013/08/veja-diferencas-entre-conceitos-que-definem-classes-sociais-no-brasil.html. Acesso em: 28 jan. 2019.

De acordo com [4], 4,4% da população sofrem de transtorno depressivo e 3,6% de transtorno de ansiedade. O Brasil apresenta a maior taxa de ansiedade do mundo. São 18.657.943 pessoas, cerca de 9,3% dos brasileiros.

As redes sociais têm ocupado um grande espaço no cotidiano das pessoas. São utilizadas para comunicação, informação, relacionamentos entre outros, e pode ser um escape para jovens que não sabem como se expressar no dia a dia. Pessoas com o perfil de ansiedade tendem a se manifestar de forma com que o outro sinta a sua dor, e possa de alguma forma ajudá-la, mas nem sempre isso acontece, gerando uma frustração maior, o que vai aumentando ainda mais os níveis de ansiedade e depressão. De acordo com a revisão bibliográfica realizada nesta pesquisa, constatou-se a inexistência de trabalhos científicos cujo tema focasse nos aspectos emocionais dos alunos dos Institutos Federais, seja no contexto da psicologia ou no contexto da computação. Dado essa necessidade, o objetivo é a criação de um modelo de aprendizado de máquina, para a geração de um dicionário intitulado como ansioso. Assim, *tweets* de pessoas relacionadas com os Institutos Federais de Educação podem ser classificados de acordo com o modelo, e o dicionário gerado poderá ser usado em outros contextos de classificações, sendo uma contribuição, visto que o mesmo não foi encontrado na literatura. A motivação para o estudo é, a criação de uma linha de pesquisa direcionada a estudantes dos Institutos Federais, de forma com que trabalhos possam ser realizados para uma conscientização a respeito de conflitos emocionais. Os resultados da análise podem servir também para que as Instituições passem a conhecer a situação dos indivíduos com os quais se relacionam, onde terão um embasamento para possíveis intervenções.

Este trabalho está organizado da seguinte forma. Na seção 1 é feita uma introdução a respeito do tema abordado neste trabalho. Na seção 2 é caracterizado de maneira geral o ambiente a ser aplicado os métodos propostos nesta pesquisa. Trabalhos correlatos encontrados na literatura são apresentados na seção 3. A seção 4 descreve e explicita os métodos utilizados no desenvolvimento desta pesquisa. Na seção 5 são apresentados os experimentos realizados, incluindo os casos de insucesso. A análise dos resultados obtidos é descrita na seção 6. Por fim, a seção 7 apresenta a conclusão e perspectivas de trabalhos futuros.

2. Contextualização do Ambiente da Pesquisa

A quantidade de dados produzida diariamente na internet, provoca mudanças na maneira pela qual as pessoas se comunicam, compartilham conhecimentos, emoções e opiniões, que influenciam o comportamento social, político e econômico em todo o mundo [5].

Segundo o relatório Digital Health [6], o número de usuários na internet ultrapassa a marca de 4 bilhões. Os usuários gastam em média, 6 horas por dia navegando na rede. Mais de 3 bilhões de pessoas ao redor do mundo

usam mídias sociais a cada mês, e, no Brasil, 78% dos usuários de internet estão em alguma rede social.

Em 2006, surgiu o *Twitter*, uma rede social com uma premissa onde cada postagem, chamada de *tweet*, não devia exceder o limite de 140 caracteres. Recentemente o *Twitter* aumentou o limite para 280 caracteres, mas a mudança não se aplica a pessoas que escrevem em japonês, chinês e coreano, pelo fato dos idiomas oferecem maior facilidade para se expressar com menos caracteres [7]. Devido à sua crescente popularidade, a rede gradativamente fez-se mais eficiente e acessível, dando aos seus usuários o poder de criar e compartilhar ideias e informações instantaneamente, sem barreiras.

A cada segundo são publicados 6.000 *tweets* no Twitter, totalizando 500 milhões de *tweets* por dia [8]. No Brasil, em 2013, a rede social também foi usada para divulgar protestos contra o aumento das tarifas de ônibus, e em 2015 e 2016 [9], para informar sobre as manifestações contra e a favor do *impeachment* da ex-presidente Dilma Rousseff. Outro movimento de grande repercussão foram as eleições presidenciais no Brasil em 2018, segundo [10]. No Brasil, 70% dos usuários do *Twitter* usam a plataforma para se informar sobre política, e mais de 60% dos usuários do *Twitter* acreditam que as ideias defendidas pelos candidatos à Presidência em seus perfis oficiais na plataforma podem contribuir com a sua decisão sobre em quem votar.

Para [11] o *Twitter* é um ótimo local para se obter opinião sobre determinado assunto, já que é uma rede social focada em opiniões de usuários. O comportamento dos usuários e suas postagens, variam de acordo com fatos temporais, como o lançamento de um novo *smartphone*, eleições presidenciais, especulações sobre a vida de artistas, entre outros. Publicações realizadas no *Twitter* se referem a acontecimentos atuais, sendo assim, uma excelente ferramenta para a coleta de dados em tempo real.

Na literatura, entre os métodos que possibilitam a extração de um conhecimento baseado na coleta, processamento e análise de dados extraídos das mídias sociais, destaca-se a análise de sentimentos, também conhecida como mineração de opinião [12]. Foi usado apenas o termo análise de sentimentos em todo o trabalho.

A análise de sentimentos trabalha em elementos do texto sem limitações de tamanho e formato, como em páginas *web*, *posts*, comentários, artigos, *tweets*, revisões de produto, entre outros. Para [13], análise de sentimentos é um conjunto de métodos, técnicas e ferramentas que visam detectar e extrair informações subjetivas, como opinião e atitudes, da linguagem. Existem pelo menos dois elementos chave em uma opinião: um alvo e um sentimento sobre este alvo [14]. Para detectar o elemento sentimento do alvo, a literatura dispõe de ferramentas que atribuem automaticamente a polaridade, com sentimentos(positivo, negativo, neutro) e emoções (satisfação, alegria, tristeza,etc.), porém, em nenhum dos métodos disponíveis o foco está em classificar a ansiedade.

3. Trabalhos Correlatos

Muitos trabalhos referentes à análise de sentimentos têm sido feitos atualmente. O trabalho [15] foi um dos primeiros com foco em classificação de *tweets*. Foi apresentado uma análise de todos os *tweets* publicados durante eventos políticos, culturais, sociais, econômicos e naturais ocorridos entre 1º de agosto e 20 de dezembro de 2008 com o objetivo de extrair e calcular seis estados de humor (tensão, depressão, raiva, vigor, fadiga, confusão) para cada dia na linha do tempo. Detectando assim que eventos no âmbito social, político, cultural e econômico têm um impacto significativo e imediato no humor público.

O trabalho [16] apresenta uma série de 21 métodos de análise em mídias sociais, a fim de identificar o método mais capaz de detectar polaridades textuais. É feita uma comparação entre estes métodos e apresentando vantagens, desvantagens e possíveis limitações de cada um. Todos os métodos estudados, apresentam duas ou três classes de classificação: positivo e negativo, ou positivo, negativo e neutro. Para o trabalho em questão, é interessante destacar o desenvolvimento de um classificador exclusivo para ansiedade.

Os autores [17] abordam a criação de um método computacional baseado em dicionário léxico, que analisa textos em português brasileiro e classifica as palavras como positivo, negativo e neutro. O método retorna os sentimentos associados com uma acurácia melhor do que os métodos disponíveis na literatura.

No trabalho de [18] é descrito a construção de uma ferramenta de coleta e análise de dados disponíveis na Internet. A análise é feita de forma probabilística com o auxílio do modelo de [19] para classificar sentimentos como, amor, ódio, vergonha, felicidade, medo entre outros.

O autor [20] detalha o desenvolvimento de um sistema, denominado AMD (Análise em Mineração de Dados), para identificar e aplicar técnicas de aprendizado de máquina em aspectos específicos em textos, utilizando a ferramenta Weka[2]. Foi utilizado uma base com polaridades (positivo e negativo) rotuladas manualmente para o treinamento do modelo.

Os trabalhos relacionados utilizam a tradução de dicionários léxicos da língua inglesa. Além disso, os rótulos disponíveis descrevem, no máximo, três grupos de sentimentos, positivo, negativo e neutro. A classificação de um texto como negativo é muito ampla, visto que, uma opinião negativa sobre um produto pode receber o mesmo rótulo que uma publicação desesperada de um suicida. A abordagem proposta neste artigo, além de trabalhar com palavras nativas da língua portuguesa, é mais acurado para sentimentos provenientes da ansiedade, pois seu treinamento se deu por escalas inerentes da psicologia. É de extrema importância ter na literatura algo para servir como base para criação de serviços ou ferramentas, que auxiliem pessoas que são acometidas

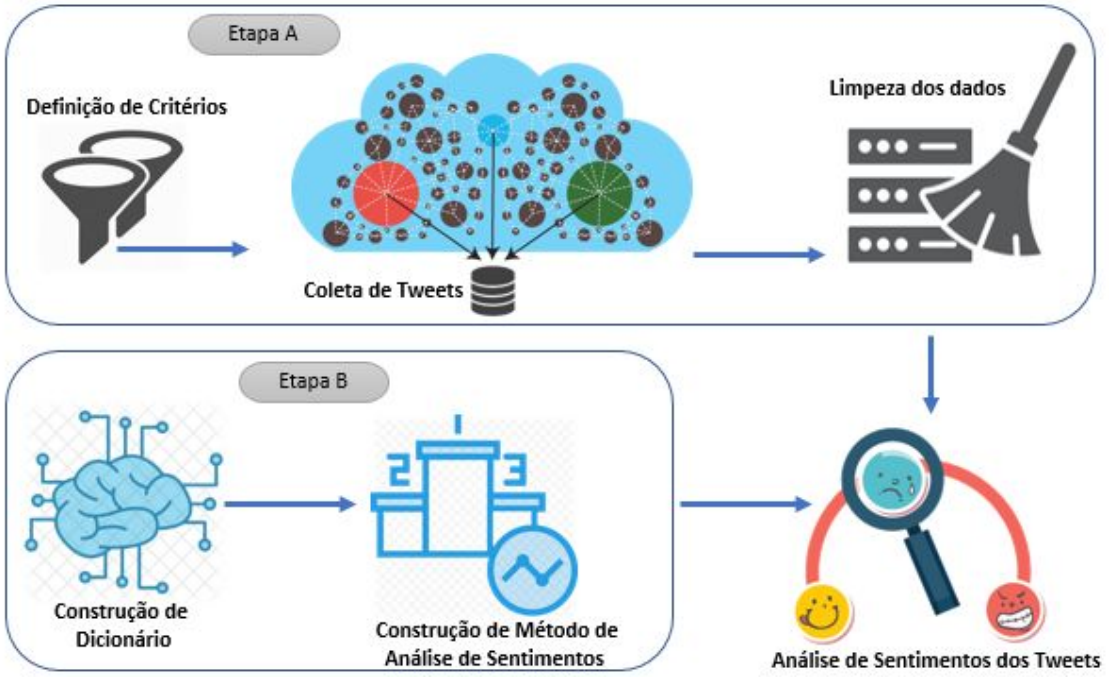
[2] UNIVERSITY OF WAIKATO. **Waikato Environment for Knowledge Analysis**. New Zealand, 2017. Disponível em: https://www.cs.waikato.ac.nz/ml/weka/downloading.html. Acesso em: 28 jan. 2019.

pela ansiedade. Principalmente no cenário em que o Brasil se encontra, com a maior porcentagem de pessoas ansiosas do mundo [6]. Outro ponto a ser destacado, é ausência de trabalhos que estudam a saúde mental de estudantes dos Institutos Federais de Educação, Ciência e Tecnologia.

4. Metodologia

O processo de identificação de perfis ansiosos consiste em um ciclo de atividades, de forma com que tenha-se um melhor aproveitamento dos dados coletados, com o menor custo computacional, visto que a base de dados requer um alto poder de processamento. Cada etapa do processo é descrito na Figura 1 e detalhado a seguir.

Figura 1. Fluxo Geral da Abordagem Proposta.



Fonte: Elaborada pelos autores

Definição de Critérios: Ao realizar buscas no *Twitter* foram identificados alguns obstáculos, como: coletar dados apenas de alunos, e, saber se a publicação (*tweet*), é de fato uma referência à Institutos Federais. Por isso, foi necessário identificar o local de estudo e o local da postagem.

- **Atributo local de estudo:** Diferente de outras redes sociais, a plataforma do *Twitter* não armazena informações referentes à atividades profissionais, como, escolaridade, local de estudo, ou trabalho. Para contornar este fato, foi adotado a estratégia de armazenar dados de todos aqueles que citaram as siglas dos Institutos Federais de Educação em suas postagens. Foi observado que a chance de analisar um perfil de usuário que não corresponde a um aluno é grande, mesmo citando alguma instituição, por exemplo: Ana é aluna. Maria não é aluna, mas é

amiga de Ana. Ana publica em seu *Twitter* que haverá um evento no IFMG. Maria visualiza a publicação de Ana, se interessa pelo evento e publica sua postagem. Conclusão: Falso positivo. Para minimizar o impacto de falso positivo, publicações republicadas foram eliminadas. Dado as circunstâncias, não é possível garantir que os *tweets* analisados pertencem somente aos alunos dos Institutos Federais, e sim afirmar que pertencem a pessoas relacionadas aos Institutos Federais.

- **Atributo TAG:** Foram usadas "tags" como filtro de coleta no Twitter. As mesmas se referem às 11 instituições federais de ensino, onde foram selecionadas de forma aleatória: "IFAC", "IFAL", "IFAM","IFAP","IFBA","IFES","IFNMG","IFMG","IFSULDEMINAS","IFMT", "IFG".
- **Atributo local de postagem:** Durante a coleta, foi observado que muitos usuários de outros países postam assuntos que contém a expressão 'IF...', trazendo irregularidades ao estudo em questão. Sendo assim, apenas postagens realizadas no Brasil, cujo idioma é Português, foram aceitas, visto que as outras não se tratavam de indivíduos com alguma ligação com os Institutos Federais, e sim de pessoas que citavam a mesma palavra, porém com outros significados.
- **Atributo idade:** O foco da pesquisa está na coleta de dados de prováveis alunos, sejam eles adolescentes, jovens ou adultos. Porém, não é possível descrever a quantidade de alunos analisados por faixa etária, visto que o Twitter não fornece a idade de seus usuários.

Coleta de *Tweets*: A extração dos dados foi possível por meio do Algoritmo I que acessa a API do *Twitter* e busca por postagens com alguma referência a Institutos Federais. Dado uma sigla x, por exemplo 'IFMG', o Algoritmo I percorre o *Twitter* em busca de postagens relacionadas. Caso encontre, o mesmo começa um processo de verificação das postagens, e só armazena se todos os critérios como idioma, localização e publicação própria forem satisfeitos. O pseudocódigo do Algoritmo I é descrito a seguir.

Início Algoritmo I

1. Seja x uma sigla referente a um Instituto Federal;
2. Leia (x);
3. **Enquanto** x estiver sendo citado **faça**
4. **Se** idioma de x == Português e localização de x == Brasil e x != retweet **faça**
5. Lista de *tweets* ← tweet e dados do usuário responsável;
6. **Fim-enquanto;**
7. Planilha recebe Lista de *tweets*;

Fim Algoritmo I

Após a execução do Algoritmo I, foram encontrados 3806 prováveis alunos, 8331 *tweets* contendo uma ou mais citações relacionadas a 32 Instituições Federais de Educação.

Também foi necessário a elaboração do Algoritmo II, para a extração de *tweets* dos usuários identificados no Algoritmo I. Através da sua execução foi possível obter 100 *tweets* do primeiro semestre de 2018, de cada um dos 3607 alunos, totalizando 360.700 *tweets*. Porém, apenas uma amostra de 100 *tweets* por aluno foi analisada, referentes a 11 Instituições de Ensino, restrito aos meses de Junho e Julho. O Algoritmo II é descrito a seguir.

Início Algoritmo II

1. Seja x um aluno de Instituição Federal;
2. Leia (x);
3. **Enquanto** x tiver dados **faça**
4. Lista de *tweets* ← tweet e dados de publicação;
5. **Fim-enquanto;**
6. Planilha recebe todos os *tweets* de aluno;

Fim Algoritmo II

Com a execução dos Algoritmos I e II , foi possível coletar dados de indivíduos que fizeram alguma postagem relacionada aos Institutos Federais. Analisar todos os *tweets* de todos os alunos se torna uma tarefa inviável, devido ao custo computacional. Por isso foi estabelecido uma amostra dentre a população para representar os alunos e suas postagens. Foi escolhido aleatoriamente uma parte da população de alunos. Assim a amostra é representada por 110 alunos, em média 100 *tweets* por pessoa, e 10 alunos por instituição. Detalhes da população e amostra dos dados coletados podem ser vistos na Tabela 1.

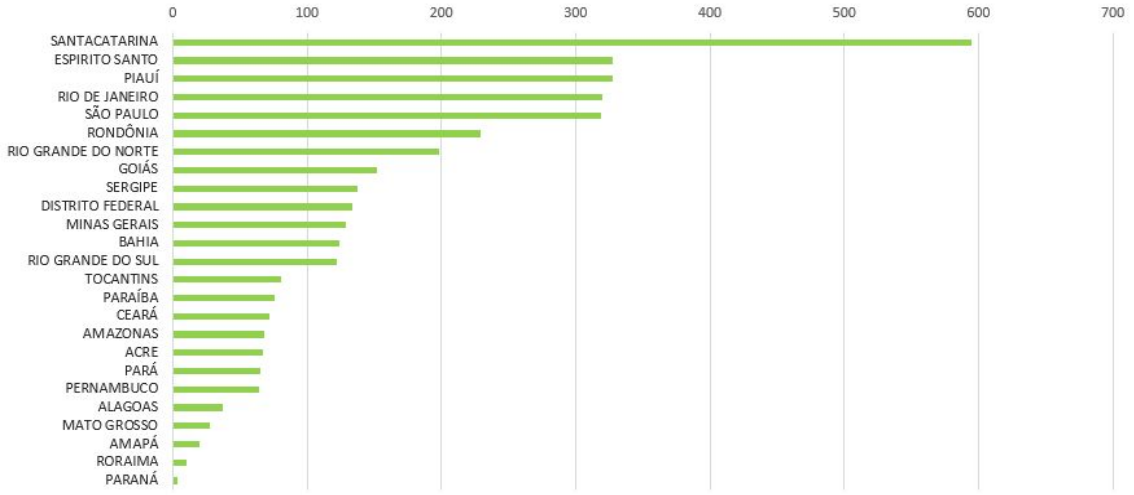
Tabela 1. Resumo dos Dados Coletados.

População Alunos	**População IF's**	**Amostra Alunos**	**Amostra IF's**	**População *Tweets***	**Amostra *Tweets***
3806	32	110	11	360.700	1.100

Fonte: Elaborada pelos autores.

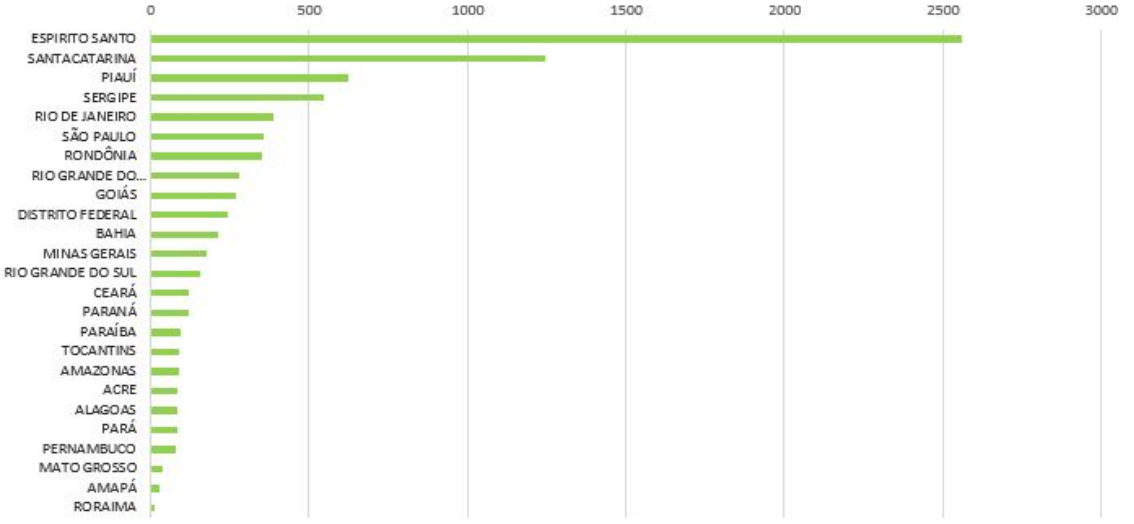
Analisando também os dados coletados, foi possível distinguir informações como quantidade de indivíduos relacionados com IF´s por estado e quantidade de *tweets* por estado como apresentado na Figura 2 e Figura 3, respectivamente.

Figura 2. Quantidade de indivíduos relacionados com IF's por Estado.



Fonte: Elaborada pelos autores

Figura 3. Quantidade de *Tweets* por Estado.



Fonte: Elaborada pelos autores.

Limpeza dos Dados: A base gerada a partir do Algoritmo I, é pré-processada e analisada. O pré-processamento dos dados é importante para a remoção de sentenças que não agregam informações relevantes ao texto, tais como, caracteres especiais, termos provenientes da plataforma, e *links*, como pode ser visto na Tabela 2.

Tabela 2. *Tweets* Coletados.

***Tweets* Originais**	***Tweets* pré-processados**
RT @id1: eu não consigo dormir direito	eu não consigo dormir direito

eu hoje =/ https://t.co/q7gUIhXoxg	eu hoje =/
@user bora estudar pra prova de amanhã?	bora estudar para a prova de amanhã?

Fonte: Elaborada pelos autores.

Construção do Dicionário: A primeira abordagem escolhida para a criação do dicionário foi a utilização de Redes Neurais [21]. Porém, o modelo gerado classificava com 98% de acurácia palavras aleatórias, como: cadeira, livro, caneta, entre outros, tornando o seu uso inviável. Assim, foi necessário adotar uma outra estratégia na construção do dicionário. A abordagem escolhida foi a aplicação de um algoritmo de modelagem de tópicos em livros de psicologia. Um exemplo de funcionamento do algoritmo pode ser visto abaixo:

Sentença: "Estou tomando café na minha xícara preferida."

Sentença: "Odeio café com leite, prefiro café puro."

Sentença: "Amo sentar na minha poltrona e tomar um café com leite."

Resultado:

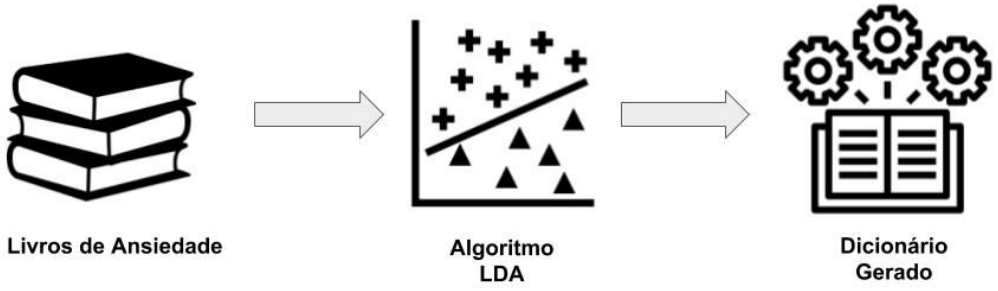
Tópico 1: "café", "leite"

Tópico 2: "xícara", "poltrona"

Tópico 3: "odeio", "amo"

A Figura 4 apresenta as etapas para a construção do dicionário, onde livros de psicologia serviram de base de dados para que o algoritmo agrupasse as palavras em tópicos. O tópico com maior frequência de palavras ansiosas é selecionado para a etapa de validação.

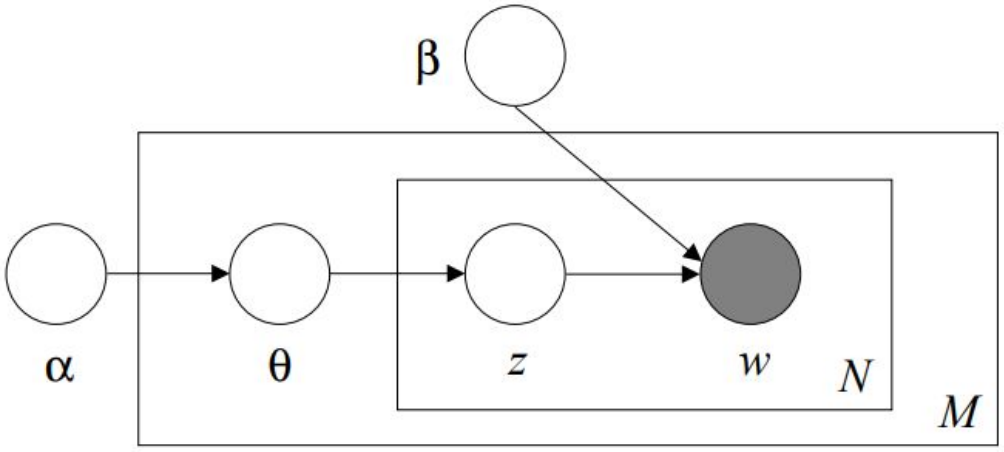
Figura 4. Fluxo de construção do dicionário.



Fonte: Elaborada pelos autores.

Construção de um Método de Análise de Sentimentos: O Método construído foi baseado no Modelo Bayesiano LDA. O LDA (Latent Dirichlet Allocation), foi inicialmente proposto no contexto da genética por [22]. Sua utilização no contexto de aprendizado de máquina foi proposto e aplicado por [23], no qual foi apresentado como um modelo gráfico para a descoberta de tópicos. Pode ser aplicado em diversos contextos, onde se têm dados não estruturados, geralmente em grande volume, onde a separação manual é inviável. Também é muito utilizado em sistemas de recomendação, como no trabalho [24], onde títulos e sinopse de filmes são analisados, para a geração de tópicos com características semelhantes. O LDA é um modelo que segue uma hierarquia de três níveis. O primeiro nível representa toda a distribuição de tópicos na coleção de documentos. No nível dois, se encontra a distribuição dos tópicos por documento. E no último nível, repete-se a distribuição dos tópicos para cada palavra em um documento. É através do último nível que se torna possível a representação de um documento como um conjunto de tópicos [25]. O funcionamento dos três níveis pode ser vistos na Figura 5.

Figura 5. Funcionamento do Modelo LDA.



Fonte: Blei e Jordan et al. (2003)

De forma a interpretar a Figura 5, [23] descreve:

> "Existem três níveis para a representação LDA. Os parâmetros α e β são parâmetros, assumidos como amostrados uma vez no processo de geração de um documento. A variável θ está no nível do documento, amostradas uma vez por documento. Finalmente, as variáveis z e w são variáveis de nível de palavra e são amostradas uma vez para cada palavra em cada documento."

As variáveis N e M indicam a quantidade de palavras, e documentos, respectivamente. É necessário observar que cada contexto exige uma configuração diferente, alguns exigirão que os parâmetros de entrada sejam definidos com um valor menor, para que a execução se dê em um tempo hábil sem uma interferência brusca no resultado.

O modelo LDA foi aplicado em uma base de dados composta por 7 livros de psicologia que retratam a ansiedade:

1. *Ansiedade - Como enfrentar o mal do século para filhos e alunos* [26].
2. *Ansiedade 3 - Ciúme* [27].
3. *Fundamentos da Psiquiatria* [28].
4. *Mentes Ansiosas* [29].
5. *Meus tempos de ansiedade: medo, esperança, terror e a busca da paz de espírito* [30].
6. *Sem Pânico* [31].
7. *Transtornos de ansiedade, estresse e depressão* [32].

Foi necessário realizar um pré-processamento na base para remoção de acentos e *stopwords*, ou seja palavras que são consideradas irrelevantes para o resultado, como: “as”, “de”, “para”, etc.

Para a remoção de *stopwords,* foi necessário a união de 5 arquivos de fontes diferentes, com palavras irrelevantes. Algumas ferramentas já fazem esse processo, porém, em um dos testes, a aplicação da biblioteca nltk nos livros não foi eficiente, visto que muitas palavras que não geram conhecimento para contexto em questão não foram removidas.

O algoritmo LDA foi responsável por gerar 5 tópicos, foi escolhido o tópico 2, onde o termo ansiedade se encontra com maior frequência. As 10 palavras com maior frequência por tópico podem ser vistas na Tabela 7.

Tabela 7. Palavras com maior frequência por tópico.

Tópico 1	Tópico 2	Tópico 3	Tópico 4	Tópico 5
pânico	ansiedade	tratamento	sintomas	corpo
medo	crises	controle	transtornos	situações
ataque	pensamentos	roberto	pacientes	crise
paciente	indivíduo	crises	sensação	maneira
transtorno	reação	silveira	médico	sensações
ataques	sintomas	família	psiquiatria	técnicas
ansiedade	casos	sente	diagnóstico	cerebral
doença	depressão	várias	normal	quadro
cérebro	corpo	medicação	sangue	normalmente

Fonte: Elaborada pelos autores.

Na Figura 6 é possível observar um mapa de distâncias interpoladas via escalonamento multidimensional [33], considerando a distribuição do tópico

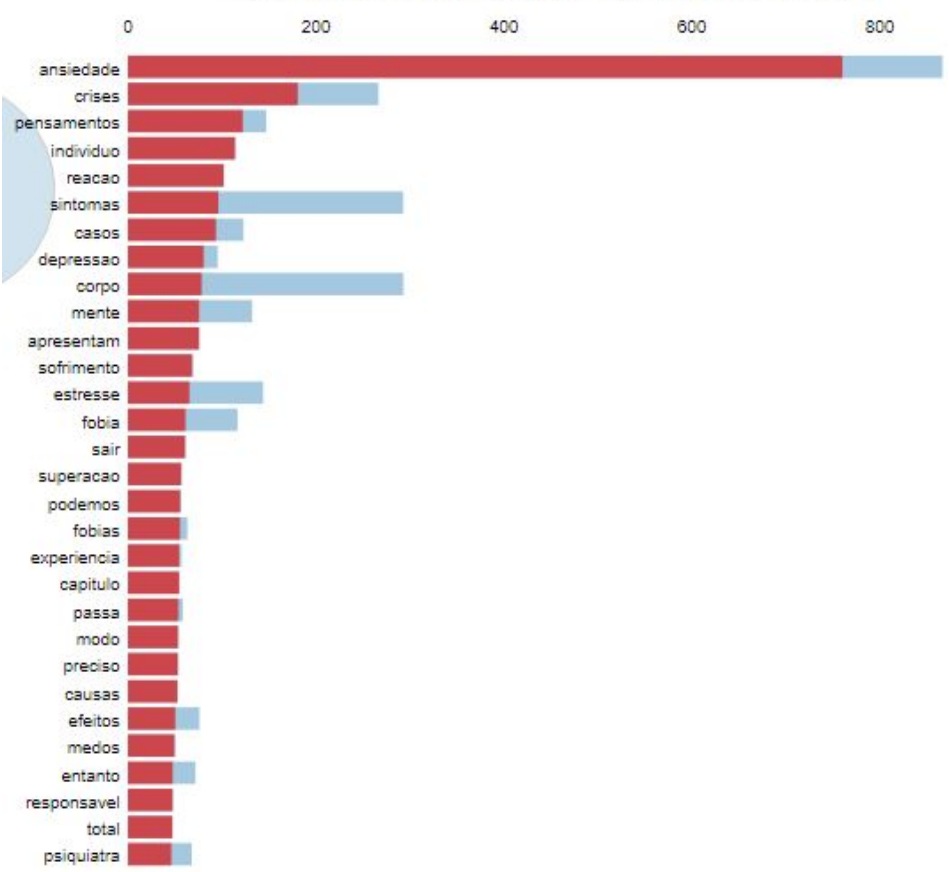
presente na margem, aplicando o primeiro e o segundo componentes principais (PC1 e PC2) [34].

Figura 6. Resultado do modelo LDA

Fonte: Elaborado pelos autores.

Na Figura 7 é exibido os 30 termos mais relevantes para o tópico 2, onde a palavra “ansiedade” é vista com maior frequência quando em comparação com outros tópicos. As barras vermelhas constituem a frequência de um termo em um tópico específico, e as barras azuis representam a frequência de um termo em todo o documento.

Figura 7. Resultado do modelo LDA



Fonte: Elaborada pelos autores.

- **Validação do Dicionário:** Após a criação do dicionário, surgiu a necessidade de validação das palavras geradas. Apesar de aplicarmos técnicas computacionais que extraem conhecimento com base em textos, os métodos não são detalhistas o suficiente para garantir a identificação de indicadores de ansiedade sem a aplicação de métodos provenientes da psicologia. Sendo assim, foi feito um agrupamento de escalas usadas na análise comportamental, e foram extraídas perguntas que serviram de validador das palavras presentes no dicionário. As escalas se referem tanto à sintomas da ansiedade quanto à sintomas da felicidade.
- **Metodologia de Validação:** As escalas de ansiedade e felicidade foram necessárias para a validação do dicionário. As mesmas foram usadas para treinar o modelo do analisador de sentimentos, assim, a máquina identificou padrões nas expressões que podem ser consideradas ansiosas e quais são tidas como expressões de felicidade. A aplicação do aprendizado de máquina nas escalas é exemplificado abaixo:

  Expressão da Escala Ansiedade: Trêmulo **Rótulo:** Ansioso

  Expressão da Escala Ansiedade: Assustado **Rótulo:** Ansioso

  Expressão da Escala Felicidade: Estou muito feliz **Rótulo:** Feliz

  Expressão da Escala Felicidade: Estou satisfeita **Rótulo:** Feliz

Na etapa de validação do dicionário foram usadas 4 escalas de ansiedade e 6 de felicidade. A descrição de cada uma delas é visto a seguir. Os detalhes podem ser encontrados no Anexo 1.

1. **Inventário de Ansiedade de Beck** [35]**:** Medida de avaliação de ansiedade mais usada, produzido para avaliar a presença e intensidade de sintomas ansiosos.
2. **Inventário de Ansiedade Traço-Estado (IDATE)** [36]**:** Apresenta uma escala que avalia a ansiedade em suas duas singularidades, a ansiedade-estado (IDATE-E) e a ansiedade-traço (IDATE-T).
3. **Escala de Ansiedade de Zung (SAS)** [37]: Questionário referente a sintomas afetivos onde, 15 sentenças expressam uma experiência negativa e 5 sentenças expressam uma experiência positiva e são pontuadas inversamente. Porém para a validação do dicionário ansioso foram usadas as 15 sentenças que apresentam sintomas negativos.
4. **Escala de Avaliação Psiquiátrica Breve (BPRS)** [38]**:** Composta por 5 perguntas que buscam avaliar o paciente e identificar o nível de ansiedade apresentado.
5. **Questionário de Felicidade de Oxford (OHQ)**[39]: Contém 29 itens para avaliar o bem-estar subjetivo. O OHQ foi derivado como um versão melhorada de seu predecessor, o Oxford Happiness Inventory [30]. Porém, dos 29 itens, apenas 17 contém referências à felicidade.
6. **Inventário de Ausência de Ansiedade (IDATE)** [40]**:** O principal objetivo da escala é a identificação de traços ansiosos, porém, o questionário contém itens associados ao fator “ansiedade ausente” e descrevem a presença de sentimentos positivos, de bem estar, satisfação, felicidade, e expressões com teor positivo.
7. **Escala de Satisfação com a vida (SWLS)** [41]: Elaborada com o objetivo de avaliar a satisfação com a vida e o bem-estar.
8. **Escala Hope** [42]**:** É uma medida de 12 itens do nível de esperança de um participante. Em particular, a escala é dividida em duas subescalas que compõem o modelo cognitivo de esperança. Dos 12 itens, 4 são compostos por elementos positivos, 4 negativos e 4 neutros, portanto foram usados apenas 4 dos 12 itens.
9. **Affectometer 2** [43]**:** Inventário de 5 minutos de felicidade geral ou sensação de bem-estar baseado na medição do equilíbrio de sentimentos positivos e negativos na experiência recente.
10. **Escala Happiness**[44]: Composto de 25 expressões, 13 referentes a depressão, e 12 afirmações que indicam o contrário. Como o objetivo nesta etapa é a construção de um modelo considerado feliz, optou-se por descartar frases referentes a depressão, usando apenas 12 expressões.

Análise de Sentimentos dos *Tweets*: De acordo com [45], a análise de sentimentos, é o campo de estudo que analisa opiniões, sentimentos, avaliações, atitudes, e principalmente a emoção em relação a produtos, serviços, organizações, indivíduos, problemas, eventos, tópicos e seus

atributos. O objetivo da análise de sentimentos é identificar e extrair de forma automática, as opiniões, sentimentos e emoções, expressados em um texto [46].

Um sentimento pode ser rotulado em classes discretas como: positivo, negativo ou neutro, ou como um intervalo que representa a intensidade deste sentimento, indicando valores acima ou abaixo de 0. Já o termo emoção é usado para intitular as percepções e pensamentos intrínsecos de uma pessoa, tais como raiva, desgosto, medo, alegria, tristeza e surpresa, não representando necessariamente um posicionamento ou uma atitude em relação ao alvo [47].

Existem diferentes níveis de análise textual, onde cada cenário exige uma granularidade diferente. Para [14], análise pode ser em nível de:

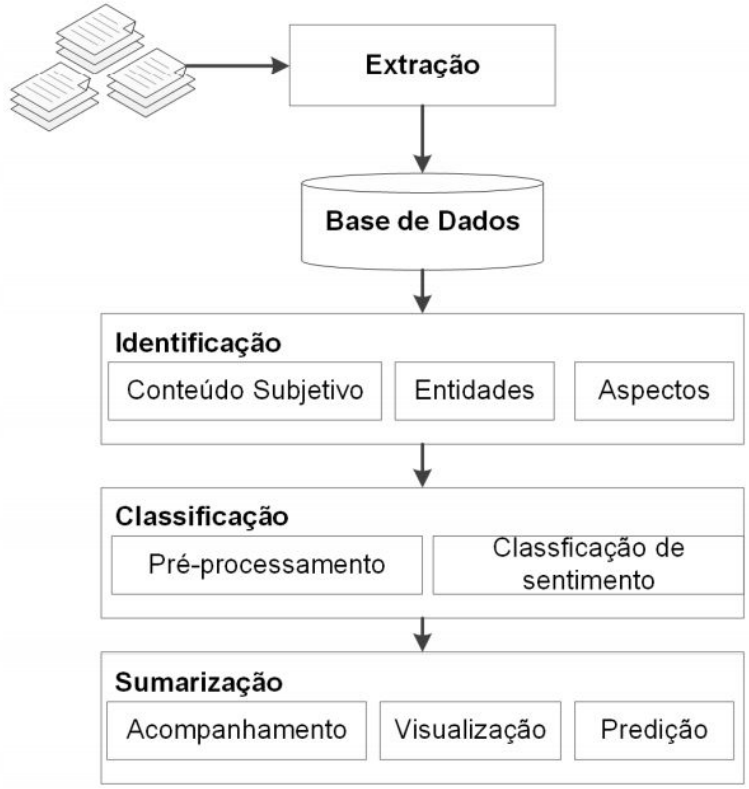
- **Documento:** Neste nível, a tarefa é classificar se uma opinião como um todo expressa um sentimento positivo ou negativo. Este nível de análise presume que cada documento expressa opiniões sobre uma única entidade, não sendo aplicável a documentos que compara várias entidades.
- **Sentença:** Determina se cada sentença expressa uma opinião positiva, negativa ou neutra. Permite identificar e distinguir sentenças objetivas (fatos) e subjetivas (opiniões). É bastante utilizado quando um mesmo documento contém opiniões sobre várias entidades.
- **Entidade e Aspecto:** Análise mais detalhada, o foco não está nas construções de linguagem (documentos, parágrafos, frases, cláusulas ou frases), e sim para a própria opinião. Por conta do aprofundamento em detalhes, este é o nível mais complexo de análise.

O processo de análise de sentimentos é composto por fases bem definidas e de igual importância, para [46] a metodologia pode ser dividida em três etapas: (1) Identificar, (2) Classificar e (3) Sumarizar. Este procedimento pode ser visualizado na Figura 9.

A **Identificação** visa encontrar tópicos existentes no conjunto de dados para uma possível associação com o conteúdo subjetivo. A complexidade da identificação muda em diferentes tipos de mídias. Documentos que se referem a uma única entidade como, opiniões sobre produtos ou serviços, tendem a ter o alvo bem definido. Já em jornais, blogs, ou posts, não se conhece os aspectos envolvidos, podendo ter várias entidades em um único trecho. A etapa de identificação envolve também o discernimento entre sentenças com ou sem opinião, eliminando trechos que não possuem qualquer relevância para o contexto [47]. Para a **Classificação** da polaridade, existem diferentes abordagens. É importante ter um bom conhecimento do que se trata a base, para a escolha do método de classificação. Em alguns cenários, os rótulos positivo e negativo conseguem englobar todas as sentenças, em outros, será necessário métodos que possuem o rótulo neutro. No entanto, classes adicionais como, muito positivo, moderado positivo, podem ser utilizadas para

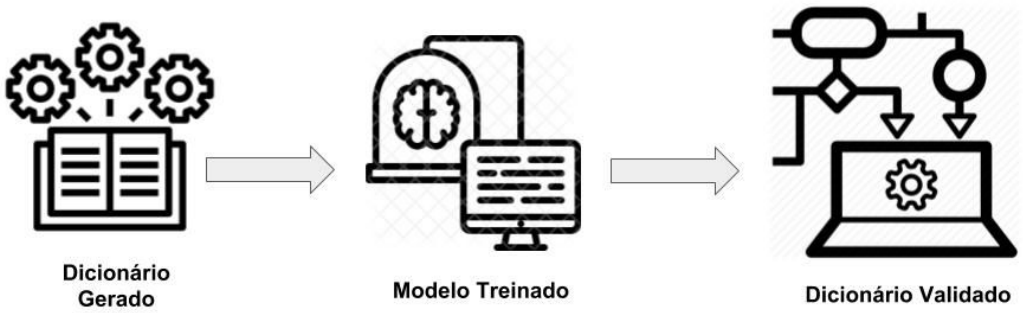
uma análise mais detalhada [46]. **Sumarização** é a etapa onde são criadas métricas que representam o sentimento de determinado tópico, como um todo, e não de forma individual. Um sumário de determinado produto ou serviço pode ser indispensável para que o usuário conheça os aspectos positivos e negativos sobre ele [46].
Para validar as palavras existentes no dicionário, o método de análise de sentimentos criado utilizou-se do algoritmo Naive Bayes, gerando um conjunto de 4.975 palavras ditas como ansiosas. O fluxo de validação do dicionário pode ser visto na Figura 9.

Figura 8. Etapas da Análise de Sentimentos.



Fonte: Tsytsarau (2012).

Figura 9. Fluxo de validação do dicionário.



Fonte: Elaborado pelos autores

O **Algoritmo Naive Bayes é** um dos algoritmos mais utilizados na análise de sentimentos [48], que aplica cálculos probabilísticos para determinar a possibilidade que o documento tem de pertencer a cada uma das categorias rotuladas. Sua representação matemática aplicada à classificação dos documentos pode ser vista no Quadro 1.

Quadro 1. Fórmula de Bayes aplicada à classificação de documentos

$$P(C = c_i \mid \vec{x}) = \frac{P(\vec{x} \mid C = c_i) \times P(C = c_i)}{P(\vec{x})}$$, **onde**:

$\vec{x}$ representa um vetor de termos e c_i representa uma classe..

Fonte: Dumais et al. (1998)

A probabilidade de cada classe pode ser facilmente encontrada levando em consideração a quantidade de documentos assimilados à classe, dividido pelo conjunto total de documentos utilizados no treinamento do classificador [49]. O algoritmo Naive Bayes pode ser utilizado para prever determinado acontecimento em tempo real, ou fazer predições de múltiplas variáveis, classificações de textos como análise de sentimentos e até mesmo sistemas de recomendação.

Os processos do algoritmo Naive Bayes foram definidos de forma bem simples por [50] e são descritos no exemplo que segue. Dado uma base rotulada, como na Tabela 3, o modelo deve classificar se a palavra “amor” é positiva ou negativa.

Tabela 3. Exemplo de base rotulada.

Palavra	Classe
Cão	Positivo
Amor	Negativo
Mau	Negativo
Amor	Positivo
Amor	Positivo
Cão	Positivo

Casa	Positivo

Fonte: Ferreira (2017)

Passo 1 : Criação de uma tabela de frequência das palavras e suas classes (Tabela 4).

Tabela 4. Frequência das palavras por classe.

Palavra	Positivo	Negativo
Cão	2	0
Amor	3	2
Mau	0	1
Gato	1	3
Casa	2	0

Fonte: Ferreira (2017)

Passo 2: Através da tabela de frequência é possível mensurar a probabilidade de ocorrência de cada termo por classe, resultando na Tabela 5.

Tabela 5. Probabilidade de ocorrência por classe.

Palavra	Positivo	Negativo	Probabilidade
Cão	2	0	2/14 = 0.14
Amor	3	2	5/14 = 0.35
Mau	0	1	1/14 = 0.072
Gato	1	3	4/14 = 0.28
Casa	2	0	2/14 = 0.14
Total	8	6	

Fonte: Ferreira (2017)

Passo 3: Calcular a probabilidade da palavra “amor” ser positiva ou negativa utilizando a fórmula do Quadro 1.

P(‘amor’|positivo) = 3/8 = 0.37 P(‘amor’|negativo) = 2/6 = 0.33

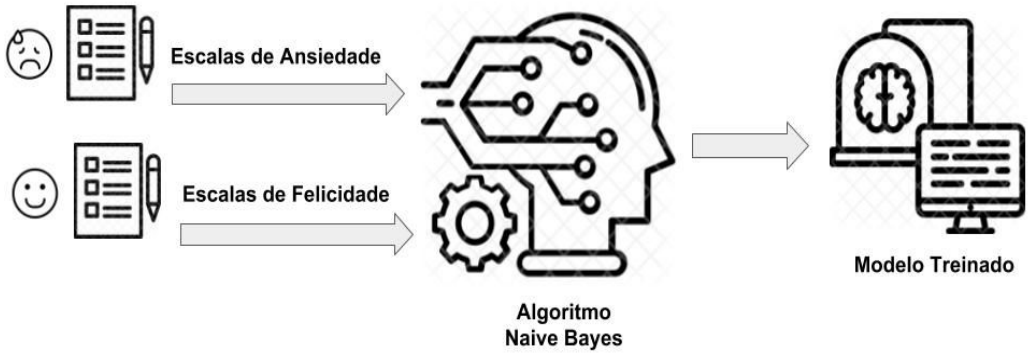
P(positivo) = 8/14 = 0.57 P(negativo) =6/14 = 0.42

P(‘amor’) = 5/14 = 0.35
P(positivo|’amor’) = 0.37 * 0.57 / 0.35 = 0.60
P(negativo|’amor’) = 0.33 * 0.42 / 0.35 = 0.39

Finalmente, a probabilidade da palavra “amor” ser positiva é maior do que a probabilidade de ser negativa. Isso ocorrerá para cada instância de teste, se a instância não for conhecida pelo modelo, a classificação é dada pela probabilidade da classe de maior frequência no treino.

A Figura 10 retrata a aplicabilidade das escalas no aprendizado de máquina.

Figura 10. Resultado do modelo LDA



Fonte: Elaborado pelos autores.

Como resultado da metodologia aplicada, foi possível conceber um algoritmo para a identificação de perfis ansiosos nos IF’s:

Início do Algoritmo para Identificação de Perfis Ansiosos nos IF's

1. Coleta de Tweets via Algoritmos I e II.
2. Limpeza dos Dados.
3. Construção do dicionário via LDA.
4. Validação do modelo via Naive Bayes usando escalas de ansiedade e felicidade.

Fim Algoritmo

5. Resultados Experimentais

A inexistência de métodos computacionais que indicam transtornos de ansiedade, ou a escassez de pesquisas que relacionam as áreas de computação e psicologia, faz com que trabalhos como este sejam baseados em experimentos. Sendo assim, foram realizados testes utilizando métodos computacionais já consolidados em uma amostra dos *tweets* pré-processados. Tanto durante a etapa de construção do dicionário, como na classificação das

sentenças, visando encontrar a forma que mais se assemelha a rotulação humana. Os resultados dos métodos apresentados nesta seção não se adequaram ao problema proposto, portanto não foram usados.

5.1. Experimentos - Classificação

Dos experimentos realizados na etapa de classificação, esperava-se encontrar o método que melhor se adequasse ao contexto ansioso. Ou seja, dado uma sentença ansiosa, a expectativa é que a mesma receba uma classificação negativa. Vários testes foram executados, mas os resultados não contemplaram a polaridade esperada. Sendo assim, se deu a necessidade de criação do método de classificação de palavras ansiosas. Os testes e resultados se encontram na seção a seguir.

5.1.1 Ferramenta iFeel

O iFeel é uma plataforma que permite polarizar sentimentos em qualquer forma de texto. Conta com 18 métodos de análise de sentimentos e seu uso é gratuito [16].

A princípio foi usado a ferramenta iFeel para a classificação de uma amostra de 500 *tweets*, usando a estratégia de votação entre os 18 métodos, a fim de obter uma maior exatidão para cada sentença. Porém, por mais que determinada expressão recebesse uma classificação negativa, não tinha como garantir o teor ansioso da mesma.

Por exemplo, o *tweet* a seguir é submetido a plataforma iFeel que classifica como positivo, negativo ou neutro:

“Se a morte bater em minha porta abrirei um sorriso e perguntarei: Porque demorou tanto?”

Pode-se observar na Figura 10 que alguns métodos rotulam a sentença como negativa, aplicando valores abaixo de 0. A dificuldade em definir uma faixa de valores aceitáveis para ansiedade reforçou o conceito de criar um método próprio para observar o teor ansioso.

Figura 10. Resultado da ferramenta iFeel.

Your input: Se a morte bater em minha porta abrirei um sorriso e perguntarei: Porque demorou tanto?

Method Name	Status	Method Score	Polarity
OPINIONLEXICON	Completed	0.33333333333333337	Positive
SENTISTRENGTH	Completed	0	Neutral
SOCAL	Completed	-0.25	Negative
HAPPINESSINDEX	Completed	-0.4075	Negative
SANN	Completed	-1	Negative
EMOTICONSDS	Completed	1	Positive
SENTIMENT140	Completed	-74.77000000000001	Negative
STANFORD	Completed	-1	Negative
AFINN	Completed	0	Neutral
MPQA	Completed	0	Neutral
NRCHASHTAG	Completed	-106.224	Negative
EMOLEX	Completed	0	Neutral
EMOTICONS	Completed	0	Neutral
PANAST	Completed	0	Neutral
SASA	Completed	-1	Negative
SENTIWORDNET	Completed	0.003277482112746243	Positive
VADER	Completed	0	Neutral
UMIGON	Completed	1	Positive

Fonte: Elaborado pelos autores.

5.1.2 LIWC

O autor [51] comparou o uso de ferramentas computacionais e a análise humana. Uma das ferramentas utilizadas na comparação é o LIWC [52] onde uma das categorias é relacionada aos níveis de ansiedade e neurose das pessoas. Porém, quando aplicada à expressões informais em tom de desabafo, a ferramenta não se mostra assertiva.

Foram feitos testes com sentenças criadas para explicitar ansiedade, e não se obteve um resultado classificado como ansioso. Os resultados sobre o teste feito, pode ser visto nas Figuras 11 e 12, utilizando as seguinte sentenças:

Teste 1: "*Meu Deus, não sei o que fazer, estudei muito e tirei 0".*

Teste 2: *"Estou ferrado, tenho prova amanhã e não consigo entender nada. "*

A ferramenta solicita dados pessoais como idade e sexo, e em que grupo a sentença de entrada se enquadra, como: mídia social, texto pessoal,comercial ou científico. Com base nas informações fornecidas, a mesma realiza uma classificação com os seguintes critérios:

1. **Auto-Referências:** Verifica a existência de palavras como: "eu", "mim", "meu", "fiz", "vou", etc.

2. **Palavras Sociais:** Analisa palavras com referência a outras pessoas.

3. **Emoções Positivas:** Identifica aspectos emotivos nas sentenças.

4. **Emoções Negativas:** Aponta palavras referenciadas como negativas.

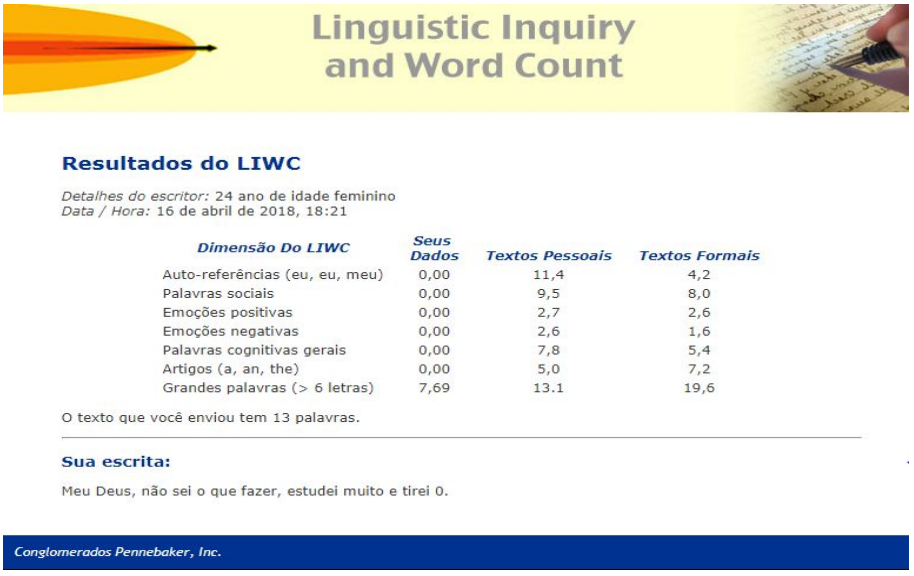
5. **Palavras Cognitivas Gerais:** Palavras que apresentam o mesmo significado.

6. **Artigos:** Identifica palavras como “a”, “o”, “uma”, “um”, etc.

7. **Grandes Palavras:** Soma de palavras com mais de 6 letras.

Além disso, o resultado apresenta duas colunas como padrão de referência, em situações de texto formal e informal.

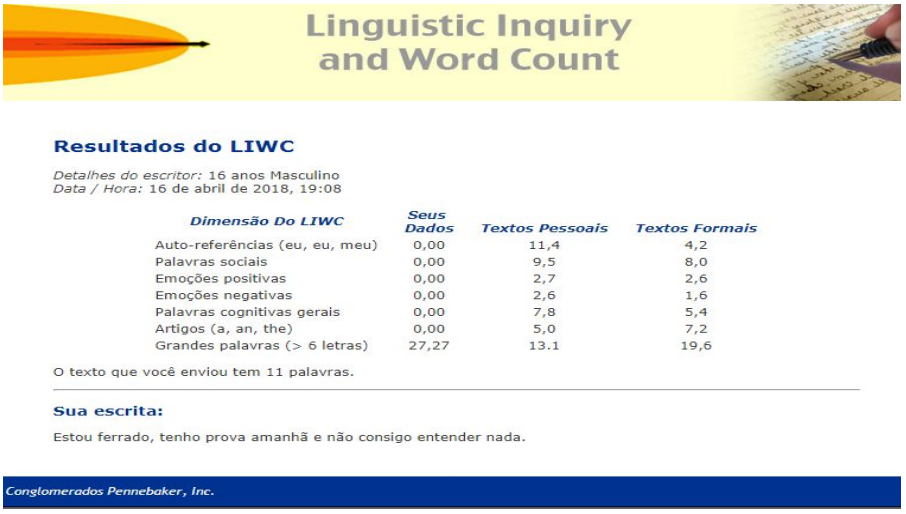
Figura 11. Resultado do teste 1.



Fonte: Elaborado pelos autores.

Nos testes realizados neste contexto, com *tweets*, foi possível observar que frases com a negatividade sutil, como um desabafo, ou um apelo, sem palavras explicitamente negativas acabaram não sendo avaliadas corretamente, sendo assim, o seu uso não oferece a precisão necessária na classificação dos *tweets*.

Figura 12. Resultado do teste 2.

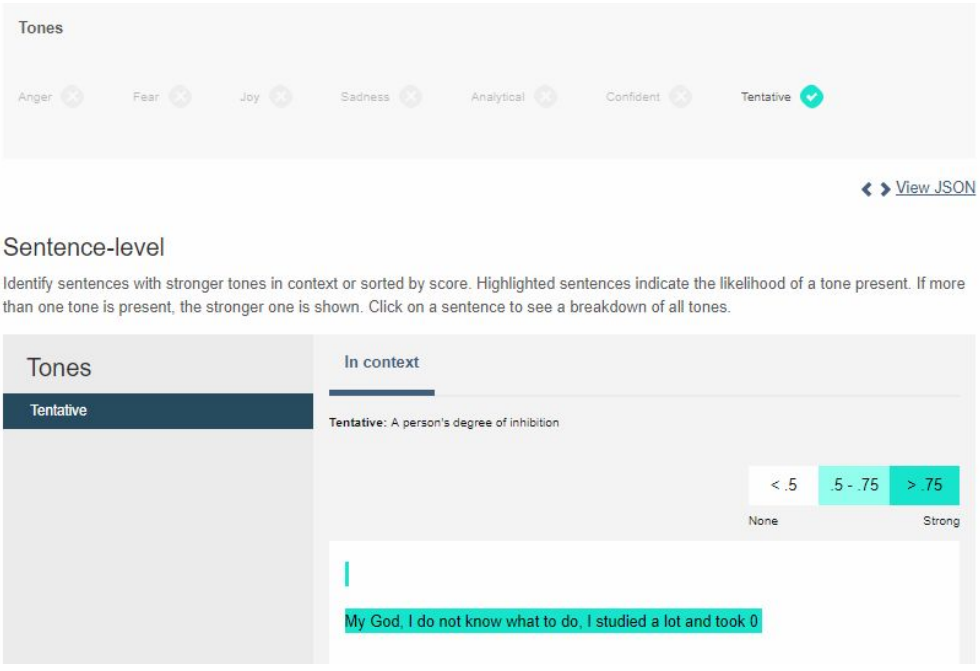


Fonte: Elaborado pelos autores.

5.1.3. IBM Watson

A ferramenta Tone Analyzer da IBM Watson [53] utiliza análise linguística para detectar alegria, medo, tristeza, raiva, tons analíticos, confiantes e hesitantes encontrados no texto. Testes foram realizados utilizando as mesmas sentenças do teste com a plataforma Liwc, porém dessa vez traduzidos para o Inglês, visto que a ferramenta Tone Analyzer não opera com expressões em Português. Os experimentos são descritos nas Figuras 13 e 14.

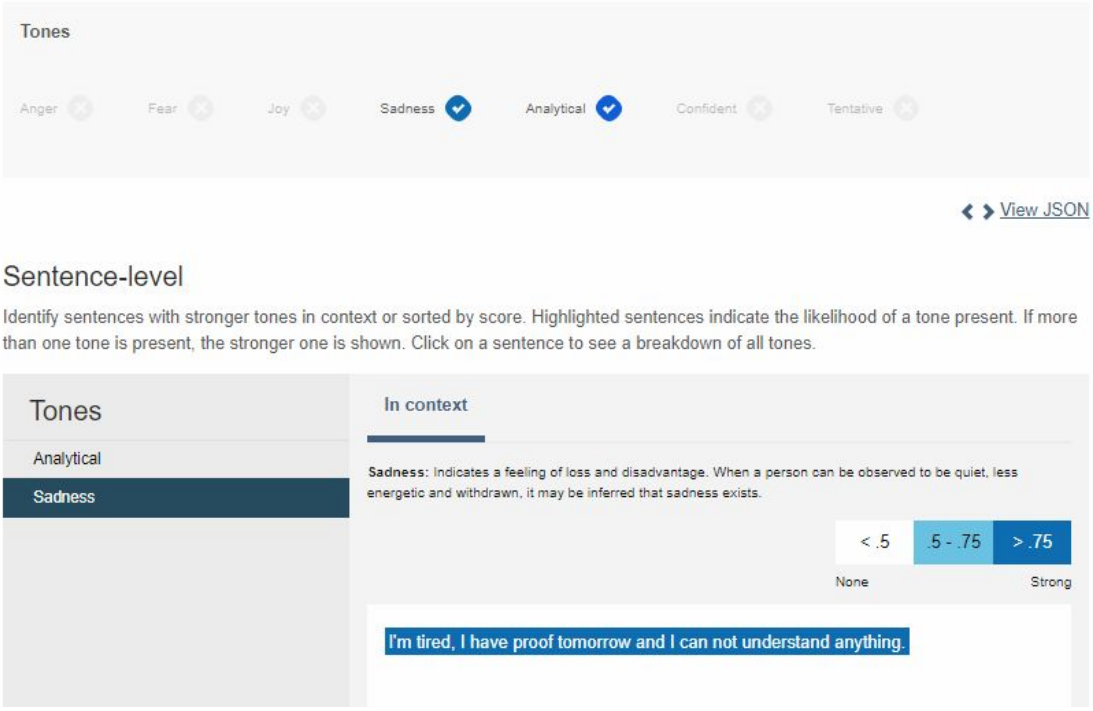
Figura 13. Resultado 1 da Ferramenta Tone Analyzer.



Fonte: Elaborado pelos autores.

A plataforma recebe uma sentença, e a atribui para os grupos rotulados com as seguintes emoções: raiva, medo, prazer, tristeza, analitico, confidente, e exitante. Apesar da existência de um grupo específico para tristeza, sentenças especificamente ansiosas não devidamente classificadas.

Figura 14. Resultado 2 da ferramenta Tone Analyzer.



Fonte: Elaborado pelos autores.

5.2. Experimentos - Dicionário

Na etapa de criação do dicionário, optou-se por utilizar o método que mais retrata o contexto ansioso, conciliando métodos da computação com os instrumentos de avaliação da psicologia. Dos métodos testados, um necessita da interferência de um profissional da área de psicologia para validar de acordo com sua experiência, o que descartaria todo o trabalho de aprendizado de máquina, e o outro teve um resultado mais simplista e dispensável para o andamento da pesquisa. Os testes e seus resultados são descritos nas seções que seguem.

5.2.1 Redes Neurais

Durante a etapa de seleção das palavras, e construção do dicionário ansioso foi realizado a implementação de uma rede neural utilizando a biblioteca Gensim do Python, que permite o treinamento de uma rede neural recorrente. Essa biblioteca recebe como entrada um conjunto de dados pré processados,

treina e apresenta como saída, palavras que mais se relacionam com a expressão ansiedade, como pode ser visto na Tabela 16.
A base de treinamento usada é a mesma do método LDA.

Tabela 16. Resultado da Rede Neural

Palavras	Porcentagem de Aproximação
medo	0.98807144
vida	0.98470723
sintomas	0.98308122
tempo	0.98192936
corpo	0.97899639
pânico	0.97802394
tratamento	0.97733185
grande	0.97673070
fobia	0.97638535
situações	0.97626250

Fonte: Elaborada pelos autores.

Ao realizar a contagem das palavras com porcentagem de aproximação maior que 80%, percebeu-se que a quantidade de palavras a serem validadas por um profissional da psicologia era inviável, totalizando 9.000 palavras. Assim, optou-se por um método que entregasse um resultado tão bom quanto, porém em menor escala.

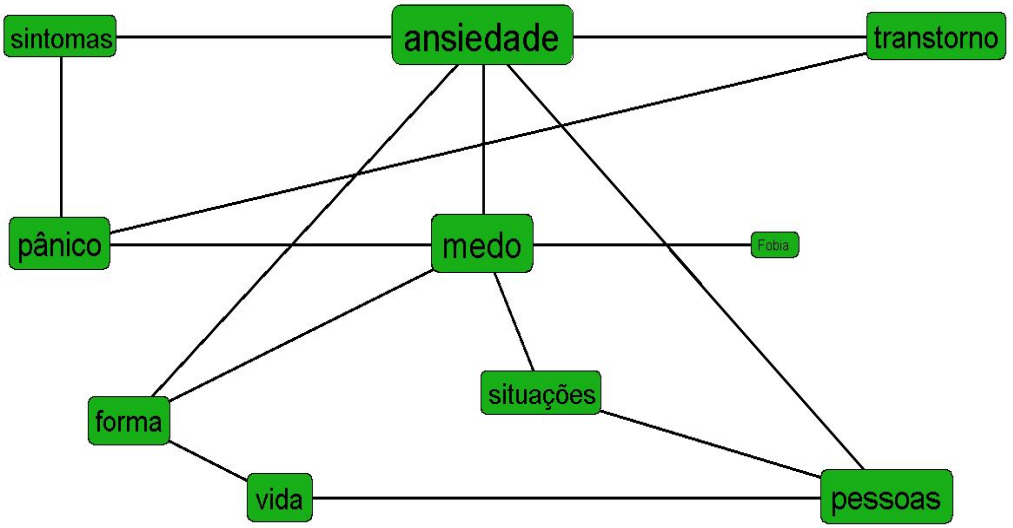
5.2.2 Sobek

Para uma visão ampla do conteúdo dos livros, foi utilizado a ferramenta de mineração Sobek[3]. Essa ferramenta foi criada em 2007 como uma ferramenta de mineração de texto para auxiliar os professores do ensino a distância a avaliarem o trabalho dos alunos. Em 2009, além de auxiliar professores, o Sobek começou a ser utilizado por alunos para a compreensão da leitura e tarefas de resumo de texto. Um ano depois, a ferramenta foi inserida em outros sistemas, tais como: avaliação de posts dos alunos em fóruns de discussão, jogos digitais para promover a narrativa escrita, ferramenta de aprendizagem baseada em projetos com recomendação de conteúdo [54].
Em uma das suas funcionalidades, é possível criar um grafo por frequência de palavras. A base com os livros ansiosos foi submetida à plataforma Sobek, onde foi pré-processada e analisada, o resultado pode ser visto na Figura 15. O

[3] UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL. **Sobek Mining**. Rio Grande do Sul, 2009. Disponível em: http://sobek.ufrgs.br/. Acesso em: 1 fev. 2019.

tamanho de cada vértice é proporcional ao número de vezes que a sentença aparece. As arestas mostram a ligação de cada palavra destacada dentro do texto.

Figura 15. Resultado da ferramenta Sobek



Fonte: Elaborado pelos autores.

Para a geração de um dicionário rotulado como ansioso, precisa-se de indicadores que vão além da frequência. Além disso, o grupo gerado precisa ter um conjunto robusto de palavras para abranger todo o contexto. Sendo assim, a ferramenta Sobek é excelente para a visualização dos dados de forma geral, porém não se aplica no desenvolvimento do dicionário em questão, visto que o dicionário precisa ser composto por palavras que abordam o mesmo tema no texto, e não somente por palavras com maior frequência.

6. Resultados

Dos experimentos realizados, optou-se por utilizar os métodos que representassem melhor o cenário ansioso/acadêmico. Sendo assim, algumas ferramentas foram desconsideradas e a criação de novos métodos foram adicionados. Obteve-se os resultados esperados, como: construção do modelo de aprendizado de máquina, geração do dicionário ansioso, e finalmente a classificação dos *tweets*, para a identificação de traços ansiosos.

6.1. Modelo Treinado

As escalas de ansiedade e felicidade foram usadas como um modelo no analisador de sentimentos. Os resultados obtidos foram satisfatórios e a taxa de acurácia foi de 81%, ou seja, a probabilidade do modelo realizar predições corretas é bem alta. Os resultados podem ser analisados na Tabela 18. Na

Tabela 19 é possível observar o desempenho do algoritmo e identificar os falsos positivos.

6.1.1 Métricas do Modelo

Existem alguns parâmetros que precisam ser observados após a construção de um modelo para mensurar a eficiência do mesmo. Dentre esses parâmetros estão:

1. **Acurácia:** Percentual de acertos do modelo.
2. **Precisão:** Número de vezes que aconteceu a predição correta de uma classe, dividido pelo número de predições da classe.
3. **Revocação:** Número de vezes que aconteceu a predição correta de uma classe, dividido pelo número de vezes que a classe aparece no teste.
4. **Medida F1:** Média entre precisão e revocação.

Tabela 18. Resultado do Analisador de Sentimentos

	Precisão	**Revocação**	**Medida F1**
0	0.91	0.71	0.80
1	0.76	0.93	0.83
Acurácia	0.8173913043478261		

Fonte: Elaborada pelos autores.

Tabela 19. Matriz de Confusão

	0	**1**	**Total**
0	41	17	58
1	4	53	57
Total	45	70	115

Fonte: Elaborada pelos autores

É possível observar na Tabela 19, que a classe 0 foi predita corretamente em 41 de 58 sentenças, errando apenas 17. A classe 1 errou apenas 4 sentenças das 57, acertando 53.

6.2. Dicionário Ansioso

Após o treino do modelo com as escalas de ansiedade e felicidade, as palavras geradas pelo modelo LDA foram submetidas para uma validação. Das 5.000 palavras, apenas 25 foram consideradas como 'felizes'. Portanto, é gerado um dicionário para a classificação de sentenças ansiosas, composto por 4.975 palavras.

Uma amostra de 20 palavras retiradas do dicionário ansioso podem ser observadas na Tabela 17. A criação do dicionário só foi possível pela união dos métodos de aprendizado de máquina, e as ferramentas da psicologia.

Tabela 17. Amostra de palavras do dicionário ansioso.

ansiedade	crises	pensamentos	sintomas	sofrimento
depressão	estresse	depressão	fobia	medos
psicoterapia	cefaleia	angústia	preocupações	trauma
obsessões	nervosismo	luta/fuga	sudorese	conflitos

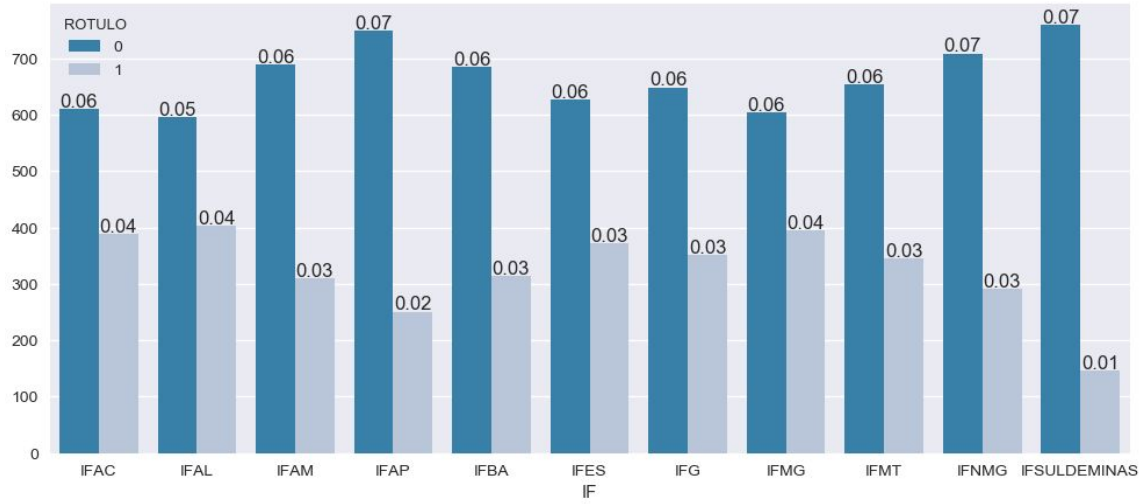
Fonte: Elaborada pelos autores.

6.3. Classificação dos *Tweets*

Com o modelo treinado, 1.100 *tweets* dos alunos foram analisados. Para evitar a desproporcionalidade entre os dados, foi considerado que cada IF possuísse a mesma quantidade de *tweets*. A análise foi feita linha a linha por instituição. Os resultados podem não refletir a realidade, visto que apenas uma amostra foi analisada. A Figura 16 apresenta um gráfico de *tweets* ansiosos e felizes por instituição analisada. O eixo x representa as instituições, enquanto o y aponta a quantidade de sentenças. O percentual de cada classe pode ser observado no topo de cada barra. O rótulo 0 indica ansiedade e o 1 felicidade.

De acordo com a análise da amostra, as instituições que mais apresentam postagens ansiosas são: IFAP e IF SUL DE MINAS, e as que se destacam com *tweets* considerados felizes são: IFAL e IFMG.

Figura 16. Resultado da análise por instituição.



Fonte: Elaborada pelos autores.

7. Conclusões

De acordo com o ponto de vista psicológico, a ansiedade é definida como um estado mental carregado de apreensão e recheado de incertezas. Se não controlada, o indivíduo tende a sofrer de outras doenças, físicas ou emocionais. Quando a ansiedade é identificada no âmbito acadêmico, fatores antes estressantes acabam sendo amenizados, visto que as Instituições envolvidas serão notificadas para que entrem com intervenções.

A coleta de dados do Twitter se deu de maneira efetiva, visto que a própria plataforma oferece meios computacionais para isso, porém é extremamente necessário a realização de um pré-processamento, visto que os dados são fornecidos com a mesma exatidão da publicação, contendo: links, hashtags, imagens e referências a outros usuários.

A utilização do algoritmo LDA foi indispensável para a extração de tópicos específicos de ansiedade dentro do conjunto de livros, mas ainda sim, foi necessário uma validação com critérios das escalas de psicologia. Sendo assim, foi possível gerar um modelo de aprendizado de máquina que classificou o conjunto de palavras gerado, como dicionário ansioso. O fato de aplicar técnicas computacionais tanto na criação, quanto validação do dicionário, dispensa qualquer interferência de viés humano.

A criação de uma base de dados exclusiva de ansiedade produz contribuições significativas na literatura, como: produção de novos estudos, classificação de dados em contextos acadêmicos ou não, e, ferramenta de auxílio para psicólogos.

8. Trabalhos Futuros

Como trabalho futuro, pretende-se realizar a classificação com base em *Hashtags* para a identificação de temas atuais mais propensos a provocar sintomas inerentes à ansiedade.

A classificação feita conteve uma amostra que não representa todos os dados coletados. Sendo assim, pretende-se reclassificar os *tweets* dos 32 IF's coletados, de forma a analisar inclusive padrões temporais. Verificando quais períodos os alunos se encontram mais ansiosos.

Apesar da utilização do algoritmo Naive Bayes para treinar o modelo com as escalas de felicidade, não foi possível gerar um dicionário de palavras felizes, portanto, espera-se produzir esse conjunto da mesma forma que foi obtido o dicionário ansioso.

Pretende-se realizar experimentos utilizando outros algoritmos de classificação além do Naive Bayes, visando obter uma melhor assertividade do modelo.

Outra vertente a ser construída é a utilização de outras abordagens no processo de manipulação dos dados além do Bag-of-words, como: Tf-Idf ou Count-Vectorizer.

Referências

[1] GENTIL, Valentim. Ansiedade e transtornos ansiosos. **Valentim Gentil, Francisco Lotufo-Neto e**, 1997.

[2] DAVIDOFF, Linda L. Introdução à psicologia. Tradução de Auriphebo Berrance Simões e Maria da Graça Lustosa. Revisão técnica de Antônio Gomes Penna. **Paulo: McGraw-Hill do Brasil**, 1983.

[3] SPIELBERGER, C. D. et al. State-trait anxiety inventory. Palo Alto. 1970.

[4] WORLD HEALTH ORGANIZATION et al. Depression and other common mental disorders: global health estimates. 2017.

[5] MONTOYO, Andrés; MARTÍNEZ-BARCO, Patricio; BALAHUR, Alexandra. Subjectivity and sentiment analysis: An overview of the current state of the area and envisaged developments. **Decision Support Systems**, v. 53, n. 4, p. 675-679, 2012.

[6] KEMP, SIMON. **DIGITAL IN 2018: WORLD'S INTERNET USERS PASS THE 4 BILLION MARK.** [*S. l.*], 2018. Disponível em: https://wearesocial.com/blog/2018/01/global-digital-report-2018. Acesso em: 2 out. 2018.

[7] G1. **Twitter aumenta limite para 280 caracteres.** [*S. l.*], 2017. Disponível em: G1. Twitter aumenta limite para 280 caracteres. [S. l.], 2017. https://g1.globo.com/tecnologia/noticia/twitter-aumenta-limite-para-280-caracteres.ghtml. Acesso em: 29 jan. 2019.

[8] PALMARELLA, Graceann et al. Advances in Digital Health Research. 2018.

[9] CORRÊA, Igor Tannús et al. Análise dos sentimentos expressos na rede social Twitter em relação aos filmes indicados ao Oscar 2017. 2017.

[10] TWITTER BRASIL. **Twitter e as #Eleições2018 no Brasil.** [S. l.], 2018. Disponível em: https://blog.twitter.com/pt_br/topics/company/2018/twitter-e-as-eleicoes-2018-no-brasil.html. Acesso em: 29 jan. 2019.

[11] RODRIGUES, Carlos Augusto S. et al. Mineração de Opinião/Análise de Sentimentos. **Trabalho acadêmico, Universidade Federal de Santa Catarina, Florianópolis. www. inf. ufsc. br/~ alvares/INE5644/MineracaoOpiniao. pdf**, 2010.

[12] NARAYANAN, Ramanathan; LIU, Bing; CHOUDHARY, Alok. Sentiment analysis of conditional sentences. In: **Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1**. Association for Computational Linguistics, 2009. p. 180-189.

[13] PANG, Bo et al. Opinion mining and sentiment analysis. **Foundations and Trends® in Information Retrieval**, v. 2, n. 1–2, p. 1-135, 2008.

[14] LIU, Bing. Sentiment analysis and opinion mining. **Synthesis lectures on human language technologies**, v. 5, n. 1, p. 1-167, 2012.

[15] BOLLEN, Johan; MAO, Huina; PEPE, Alberto. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. **Icwsm**, v. 11, p. 450-453, 2011.

[16] BENEVENUTO, Fabrício; RIBEIRO, Filipe; ARAÚJO, Matheus. Métodos para análise de sentimentos em mídias sociais. 2015.

[17] DE SOUZA, Karine França; PEREIRA, Moisés Henrique Ramos; DALIP, Daniel Hasan. UniLex: Método Léxico para Análise de Sentimentos Textuais sobre Conteúdo de Tweets em Português Brasileiro. **Abakós**, v. 5, n. 2, p. 79-96, 2017.

[18] DA SILVA, Rafael Silva. SISTEMA PARA IDENTIFICAÇÃO DE SENTIMENTOS EM TEXTOS NA WEB.

[19] ORTONY, Andrew; CLORE, Gerald L.; COLLINS, Allan. **The cognitive structure of emotions**. Cambridge university press, 1990.

[20] SAUSEN, Frederico Jacobi. **Projeto e desenvolvimento de um sistema para definição de aspectos e análise de sentimentos em textos**. 2016. Trabalho de Conclusão de Curso.

[21] MCCULLOCH, Warren S.; PITTS, Walter. A logical calculus of the ideas immanent in nervous activity. **The bulletin of mathematical biophysics**, v. 5, n. 4, p. 115-133, 1943.

[22] PRITCHARD, Jonathan K.; STEPHENS, Matthew; DONNELLY, Peter. Inference of population structure using multilocus genotype data. **Genetics**, v. 155, n. 2, p. 945-959, 2000.

[23] BLEI, David M.; NG, Andrew Y.; JORDAN, Michael I. Latent dirichlet allocation. **Journal of machine Learning research**, v. 3, n. Jan, p. 993-1022, 2003.

[24] RODRIGUES, O. A. M. ; LACERDA, A. M. ; PADUA, F. L. C. . Improved Cold-Start Recommendation via Two-Level Bandit Algorithms. In: XIV National Meeting on Artificial and Computational Intelligence, 2017, Uberlândia. Anais do XIV ENIAC, 2017. v. 1. p. 1-12.

[25] FALEIROS, Thiago de Paulo et al. Modelos probabilísticos de tópicos: desvendando o latent Dirichlet allocation. 2016.

[26] CURY, AUGUSTO JORGE. **ANSIEDADE-Como enfrentar o mal do século para filhos e alunos**. Editora Saraiva, 2017.

[27] CURY, Augusto. **Ansiedade 3. Ciúme**. Editora Saraiva, 2017.

[28] SILVEIRA, Roberto. **Fundamentos da Psiquiatria**. Editora Lumen Juris, 2014.

[29] SILVA, Ana Beatriz Barbosa. **Mentes ansiosas: medo e ansiedade além dos limites**. Editora Objetiva, 2011.

[30] STOSSEL, Scott. **Meus tempos de ansiedade: medo, esperança, terror e a busca da paz de espírito**. Editora Companhia das Letras, 2014.

[31]BARRY, Joseph. **Sem pânico**. E-book, 2011.

[32] SERSON, Breno. **TRANSTORNOS DE ANSIEDADE, ESTRESSE E DEPRESSÕES: Conhecer e tratar**. MG Editores, 2016.

[33] HERDEIRO, Roberto Francisco Casagrande. Escalonamento multidimensional. **Análise Multivariada. São Paulo: Atlas**, 2007.

[34] MARTÍNEZ, Aleix M.; KAK, Avinash C. Pca versus lda. **IEEE Transactions on Pattern Analysis & Machine Intelligence**, n. 2, p. 228-233, 2001.

[35] GORENSTEIN, C.; ANDRADE, L. H. S. G.; ZUARDI, A. Beck Depression Inventory: psychometric properties of the Portuguese version. **Rev Psiq Clin**, v. 25, n. 5, p. 245-250, 1998.

[36] SPIELBERGER, Richard L. Gorsuch. State-Trait Anxiety Inventory. Charles D. 1970.

[37] ZUNG, William W. A rating instrument for anxiety disorders. **Psychosomatics: Journal of Consultation and Liaison Psychiatry**, 1971.

[38] OVERALL, John E.; GORHAM, Donald R. The brief psychiatric rating scale. **Psychological reports**, v. 10, n. 3, p. 799-812, 1962.ar

[39] HILLS, Peter; ARGYLE, Michael. The Oxford Happiness Questionnaire: a compact scale for the measurement of psychological well-being. **Personality and individual differences**, v. 33, n. 7, p. 1073-1082, 2002.

[40] ARGYLE, Michael; MARTIN, Maryanne; CROSSLAND, Jill. Happiness as a function of personality and social encounters. **Recent advances in social psychology: An international perspective**, p. 189-203, 1989.

[41] DIENER, E. D. et al. The satisfaction with life scale. **Journal of personality assessment**, v. 49, n. 1, p. 71-75, 1985.

[42] BABYAK, Michael A.; SNYDER, C. R.; YOSHINOBU, Lauren. Psychometric properties of the Hope Scale: A confirmatory factor analysis. **Journal of Research in Personality**, v. 27, n. 2, p. 154-169, 1993.

[43] KAMMANN, Richard; FLETT, Ross. Affectometer 2: A scale to measure current level of general happiness. **Australian journal of psychology**, v. 35, n. 2, p. 259-265, 1983.

[44] MCGREAL, Rita; JOSEPH, Stephen. The depression-happiness scale. **Psychological Reports**, v. 73, n. 3_suppl, p. 1279-1282, 1993.

[45] LIU, Bing; ZHANG, Lei. A survey of opinion mining and sentiment analysis. In: **Mining text data**. Springer, Boston, MA, 2012. p. 415-463.

[46] TSYTSARAU, Mikalai; PALPANAS, Themis. Survey on mining subjective data on the web. **Data Mining and Knowledge Discovery**, v. 24, n. 3, p. 478-514, 2012.

[47] BECKER, Karin; TUMITAN, Diego. Introdução à mineração de opiniões: Conceitos, aplicações e desafios. **Simpósio brasileiro de banco de dados**, v. 75, 2013.

[48] DUMAIS, Susan et al. Inductive learning algorithms and representations for text categorization. In: **Proceedings of the seventh international conference on Information and knowledge management**. ACM, 1998. p. 148-155.

[49] MITCHELL, Tom M. et al. Machine learning. WCB. 1997.

[50] FERREIRA, Rodrigo. **Análise de sentimentos – Aprenda de uma vez por todas como funciona utilizando dados do Twitter**. [S. l.], 2017. Disponível em: https://imasters.com.br/desenvolvimento/analise-de-sentimentos-aprenda-de-uma-vez-por-todas-como-funciona-utilizando-dados-do-twitter. Acesso em: 2 fev. 2019.

[51] DE SALES MOREIRA, Vanessa et al. Análise de Sentimentos: Comparando o uso de ferramentas e a análise humana. 2016.

[52] Linguistic Inquire and Word Count. Disponível em: <www.liwc.net/>. Acesso em 16, abril, 2018.

[53] IBM. Watson. [S. l.], 2010. Disponível em: https://www.ibm.com/watson/services/tone-analyzer/. Acesso em: 20 nov. 2018.

[54] REATEGUI, Eliseo et al. Sobek: A text mining tool for educational applications. In: **Proceedings of the International Conference on Data Mining (DMIN)**. The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp), 2011. p. 1.

ANEXO I

Tabela 1. Inventário de Ansiedade de Beck.

Dormência ou formigamento	Sem equilíbrio	Dificuldade de respirar
Sensação de calor	Aterrorizado	Medo de morrer
Tremores nas pernas	Nervoso	Assustado
Incapaz de relaxar	Sensação de sufocação	Indigestão ou desconforto no abdômen
Medo que aconteça o pior	Tremores nas mãos	Sensação de desmaio
Atordoado ou tonto	Trêmulo	Rosto afogueado
Palpitação ou aceleração do coração	Medo de perder o controle	Suor

Fonte: Gorenstein et al. (1998)

Tabela 2. Escala de Ansiedade Traço do IDATE.

Idéias sem importância me preocupam	Tenho vontade chorar
Dificuldade em tomar decisões	Deixo me afetar muito pelas coisas
Dificuldades se acumulando	Não tenho confiança em mim
Levo desapontamentos a sério	Canso-me facilmente
Tenso(a) e perturbado(a) com problemas	Queria ser tão feliz quanto os outros
Sinto-me deprimido(a)	Evito dificuldades
Preocupo-me com coisas sem importância	

Fonte: Spielberger (1970)

Tabela 5. Zung self-rating anxiety scales (SAS).

Sinto-me mais nervoso e ansioso do que costume	Sinto o meu coração a bater depressa demais
Sinto-me com medo sem nenhuma razão para isso	Tenho crises de tonturas que me incomodam
Sinto-me facilmente perturbado ou em pânico	Tenho crises de desmaio ou a sensação de que vou desmaiar
Sinto-me como se estivesse para “rebentar”	Sinto os dedos das minhas mãos e dos meus pés entorpecidos e com picadas

Tenho pesadelos	Costumo ter dores de estômago ou más digestões
Sinto os braços e as pernas a tremer	Tenho de esvaziar a bexiga com frequência
Tenho dores de cabeça, no pescoço e nas costas, que me incomodam	A minha face costuma ficar quente e corada
Sinto-me fraco e fico facilmente cansado	

Fonte: Zung (1971)

Tabela 6. Escala de Avaliação Psiquiátrica Breve(BPRS)

Você está preocupado com alguma coisa?
Você tem se sentido tenso ou ansioso a maior parte do tempo?
Quando se sente assim, você consegue saber o porquê?
De que forma suas ansiedades ou preocupações afetam o seu dia a dia?
Existe algo que ajuda a melhorar essa sensação?

Fonte: Overall e Gorham (1962)

Tabela 7. The Oxford Happiness Questionnaire.

Eu estou intensamente interessado em outras pessoas	Eu tenho sentimentos muito calorosos em relação a quase todos
Eu sinto que a vida é muito gratificante	Estou bem satisfeito com tudo na minha vida
Eu ri muito	Estou muito feliz
Eu acho a maioria das coisas divertidas	Eu acho beleza em algumas coisas
Estou sempre comprometida e envolvida	Eu sempre tenho um efeito alegre nos outros
A vida é boa	Eu posso encaixar em tudo que eu quero
Eu me sinto capaz de ter qualquer coisa	Eu freqüentemente sinto alegria e euforia
Eu me sinto totalmente mentalmente alerta	Eu sinto que tenho muita energia
Eu geralmente tenho uma boa influência em eventos	

Fonte: Hills e Peter (2002)

Tabela 8. Inventário de Ansiedade Ausente de IDATE.

Estou satisfeito(a)	Sinto-me seguro(a)

Sou estável	Sou calmo(a), ponderado(a)
Sinto-me seguro	Sinto-me descansado(a)
Estou satisfeita	

Fonte: Spielberger (1970)

Tabela 9. escala Satisfaction With Life Scale (SWLS)

De muitas maneiras minha vida é perto do meu ideal.
As condições da minha vida são excelentes.
Até agora eu tenho conseguido as coisas importantes que eu quero em vida.
Se eu pudesse viver minha vida.
Eu mudaria quase nada.

Fonte: Diener et al. (1985)

Tabela 10. Hope Scale

Eu energicamente busco meus objetivos
Minhas experiências passadas me prepararam bem para o meu futuro
Eu tenho uma vida muito bem sucedida
Eu conheço as metas que estabeleci para mim

Fonte: Babyak et al. (1983)

Tabela 11. Affectometer 2

Minha vida está no caminho certo.	Eu posso lidar com todos os problemas que surgem.
Meu futuro parece bom.	Eu me sinto perto das pessoas ao meu redor.
Eu gosto de mim mesmo.	Eu tenho energia de sobra.
Eu me sinto amada e confiável.	Eu sorrio e rio muito.
Eu penso claramente e criativamente	

Fonte: Kammann e Flett (1983)

Tabela 12. Depression–Happiness Scale (D–H S)

Eu me sinto mentalmente alerta	Eu me senti feliz
Eu me senti alegre	Eu me senti otimista sobre o futuro

Eu me senti satisfeito com a minha vida	Eu senti que a vida era gratificante
Eu me senti saudável	Eu senti que a vida tinha um propósito
Eu senti que tinha sido bem sucedido	Eu me senti satisfeito com o jeito que sou
Eu achei fácil tomar decisões	Eu senti que a vida era agradável

Fonte: Mcgeral e Joseph (1993)