

SentimentalPeaks: Um modelo de aprendizado de máquina para classificação e visualização de sentimentos em tweets

Gustavo Hernandez Duarte, Isabel Harb Manssour

Escola Politécnica

Pontifícia Universidade Católica do Rio Grande do Sul (PUCRS)

Porto Alegre — RS — Brasil

Resumo—Nos dias atuais a rede social Twitter vem sendo amplamente utilizada para se interagir com transmissões televisivas ao vivo, permitindo que os telespectadores manifestem suas opiniões. O entendimento destes dados pode ser de grande importância para compreender o que os usuários pensam a respeito de um determinado tema. Portanto, desenvolvemos um modelo de aprendizado de máquina que combina 3 algoritmos, *Support Vector Machine*, Regressão Logística e *Naive Bayes*, para classificar emoções em tweets em português de diferentes assuntos. Nosso modelo obteve uma acurácia de 71% na classificação de emoções em positivo, negativo e neutro. Também integramos este modelo ao PeakVis, uma ferramenta interativa que possibilita sincronizar uma gravação de vídeo a um conjunto de tweets, incluindo uma visualização dos tweets classificados. Com esta visualização podemos aplicar filtros e analisar como um grupo de usuários se sente, por exemplo, em relação a algum jogo de futebol, debate político, capítulo de novela ou *reality show*.

Palavras chaves—Aprendizado de máquina, Análise de sentimento, Visualização de dados, Twitter

I. INTRODUÇÃO

Nos dias atuais temos uma quantidade massiva de dados sendo geradas por usuários na internet [20], especialmente quando falamos de redes sociais, através das quais os usuários expressam suas opiniões sobre diversos assuntos. Esses dados podem carregar algum sentimento que pode ser de grande valor para empresas que desejam entender como o público enxerga sua marca ou produto, para profissionais de outras áreas como Ciência de Dados que desejam obter *insights* e análises baseados nos sentimentos, e até mesmo para jornalistas que querem ter uma percepção do sentimento das pessoas em relação a uma celebridade ou evento.

Processamento de Linguagem Natural (PLN) é uma área da Ciência da Computação que estuda o desenvolvimento de programas de computador que analisam, reconhecem e/ou geram textos em linguagens humanas, ou linguagens naturais [15]. Análise de sentimento, ou mineração de opinião, é uma subárea de PLN que visa identificar e extraír opiniões dentro de um determinado texto [12]. O seu principal objetivo é identificar os sentimentos e as emoções de um indivíduo com base no tratamento computacional da subjetividade em um texto [12]. Diversos algoritmos de aprendizado de máquina são utilizados para automatizar o processo de análise de sentimento [22]. A ideia é identificar o sentimento dos usuários, ou verificar a sua opinião a respeito de um determinado tema,

tal como uma eleição presidencial, através de suas postagens em redes sociais [7].

O Twitter é uma rede social que possui o estilo de microblog, na qual os usuários postam textos, de no máximo 280 caracteres. Ele é bastante utilizado para análise de sentimentos devido a quantidade de dados e sua facilidade em extrair dados através de uma API própria fornecida para esta finalidade, pois os tweets são considerados públicos conforme a política de privacidade da plataforma [4]. Por isso, por exemplo, o Twitter tem sido utilizado para analisar as opiniões dos usuários sobre um determinado produto [22] ou sobre algum candidato a presidência em uma eleição [7].

O Twitter também tem sido utilizado como uma nova forma de interação com transmissões televisivas, quando uma transmissão de televisão como novelas, reality shows e esportes é acompanhada por um fluxo de postagens feita pelo público. Assim, consegue-se mediar a relação assimétrica entre a rede de televisão e os espectadores por meio de comentários, humor e crítica [14]. Neste contexto, foi desenvolvida a ferramenta PeakVis com o objetivo de ajudar a entender essas relações [14]. PeakVis¹ é uma ferramenta interativa que possibilita sincronizar uma gravação de vídeo a um gráfico de linha gerado a partir de um conjunto de dados de mídia social, no caso, o Twitter. Assim, permite selecionar e analisar os destaques da transmissão com base nos picos de tráfego do Twitter, e também visualizar dinamicamente as mensagens mais retuitadas e as palavras mais conectadas [14].

Portanto, visando auxiliar na análise de sentimentos, foi desenvolvido neste trabalho um modelo de aprendizado de máquina capaz de classificar sentimentos de tweets em português de diversos assuntos. Para isto, foram combinados três algoritmos de aprendizado de máquina, *Support Vector Machine*, Regressão Logística e *Naive Bayes*, obtendo uma acurácia de 71%. Estes algoritmos foram treinados com um conjunto de dados de 3700 tweets coletados sobre diferentes assuntos, anotados manualmente por três anotadores e pré-processados. Também foi desenvolvida uma visualização para os tweets classificados que foi integrada à ferramenta PeakVis. Esta combinação do modelo de aprendizado de máquina com as visualizações chamamos de SentimentalPeaks.

O restante deste trabalho está organizado da seguinte forma:

¹<https://github.com/DAVINTLAB/Peakvis>

a Seção 2 apresenta a fundamentação teórica do nosso trabalho, listando os algoritmos que foram utilizados e métricas de validação. A Seção 3 apresenta alguns trabalhos anteriores que serviram de base para o nosso trabalho. A Seção 4 apresenta a metodologia do trabalho, explicando todas as etapas realizadas. A Seção 5 descreve de forma detalhada o conjunto de dados utilizado. Na Seção 6 apresentamos o treinamento dos algoritmos de aprendizado de máquina e o nosso modelo desenvolvimento. A visualização de sentimentos dos tweets é apresentada na seção 7, os resultados obtidos e um estudo de caso são descritos na seção 8. Por fim, a Seção 9 contém as conclusões e trabalhos futuros.

II. FUNDAMENTAÇÃO TEÓRICA

A utilização de aprendizado de máquina é uma das principais maneiras de se realizar análise de sentimentos em tweets [12]. Para isso, diversas abordagens e algoritmos podem ser utilizados [12]. Nas seções a seguir explicaremos a abordagem escolhida e os algoritmos que foram utilizados no nosso trabalho.

A. Análise de sentimento

A análise de sentimentos é realizada por meio de algoritmos de PLN que realizam a análise de texto utilizando recursos computacionais, tendo a finalidade de criar conhecimento a partir destes dados [12]. Existem duas abordagens conhecidas para análise de sentimento: léxica e aprendizado de máquina. Na abordagem léxica a classificação do sentimento depende de um dicionário contendo palavras classificadas em cada classe (positivo, negativo e neutro). O texto, então, é classificado considerando a quantidade de palavras em cada classe [8]. A segunda abordagem, aprendizado de máquina, consiste em utilizar algoritmos que geram programas ou funções matemáticas que classificam sentimentos sem necessitar de um dicionário de palavras como na abordagem léxica [13]. Neste trabalho será utilizada a abordagem de aprendizado de máquina.

B. Aprendizado de máquina

Aprendizado de máquina é uma área da inteligência artificial na qual computadores são programados para aprender com a experiência passada. Para isso empregam um princípio de inferência denominado indução, no qual se obtém conclusões genéricas a partir de um conjunto particular de exemplos [6].

O aprendizado de máquina é dividido em paradigmas de aprendizados, sendo eles preditivo ou aprendizado supervisionado, e descritivo ou aprendizado não supervisionado [6]. No aprendizado supervisionado o objetivo é encontrar uma hipótese, também chamada de função, a partir dos dados de treinamento que possa ser utilizada para prever um rótulo ou valor que caracteriza um novo exemplo, com base nos valores de seus atributos de entrada. Para isso, cada objeto do conjunto de treinamento deve possuir atributos de entrada e saída [6].

Já no aprendizado não supervisionado o objetivo é explorar ou descrever um conjunto de dados. Neste tipo de aprendizado o objetivo é aprender com dados de teste que não foram previamente rotulados ou classificados, identificando semelhanças

nos dados e reagindo com base na presença ou ausência dessa semelhanças. Os algoritmos utilizados nesse paradigma não usam o atributo de saída do conjunto de dados [6].

Algoritmos de aprendizado de máquina aprendem a induzir uma função ou hipótese capaz de resolver um problema a partir de dados que representam instâncias do problema a ser resolvido [6]. Cada algoritmo utiliza uma forma para descrever a hipótese indutiva. Nossa trabalho utiliza o paradigma de aprendizado supervisionado, com os seguintes algoritmos:

- **Régressão Logística:** Tem como objetivo produzir, a partir de um conjunto de observações, um modelo que permita a predição de valores para uma variável categórica, frequentemente binária ou seja, (S/N), (1/0) ou (Verdadeiro/Falso), utilizando estatística [20]. Por exemplo, poderíamos utilizar este algoritmo para classificar se um e-mail é *spam* ou não, situação na qual há uma predição categórica binária;
- **Multinomial Naive Bayes:** É um algoritmo que utiliza a probabilidade e o teorema de Bayes para realizar predições, ou seja, todas as variáveis são consideradas independentes umas das outras [20]. Seu funcionamento consiste em verificar a probabilidade de uma variável pertencer a uma determinada categoria [20]. Por exemplo, este algoritmo pode ser utilizado para classificação de sentimentos em texto. Neste caso, ao classificar um texto o algoritmo verificaria a probabilidade de cada palavra pertencer a alguma categoria, e com base na maior probabilidade encontrada com a soma das probabilidades de cada palavra, informaria a qual categoria o texto pertence;
- **Support vector machine:** É um algoritmo não probabilístico, no qual, considerando um conjunto de dados de exemplo para treinamento, sendo cada dado marcado como pertencente a alguma categoria, ele constrói um modelo que mapeia cada dado para pontos no espaço. Novos exemplos são mapeados nesse mesmo espaço e considerados pertencentes a uma categoria com base em qual posição neste espaço eles se enquadram [20]. Poderíamos utilizar este algoritmo, por exemplo, para classificação de espécies de animais em conjunto de dados pequeno, no qual o algoritmo mapearia cada dado em um ponto no espaço não consumindo muita memória, e a cada novo dado para classificar, verificaria em qual ponto esse dado estaria, e, a partir disto, informaria a qual espécie este dado pertence.

C. Métricas de Avaliação

Para entender se um modelo de aprendizado de máquina supervisionado tem resultados considerados bons ou ruins existem algumas métricas de avaliação [6] sendo elas:

- **Matriz de Confusão:** Representa a frequência de classificação para cada uma das categorias do problema [10]. Esta métrica mostra como são separados os valores de Verdadeiros Positivos (VP), Verdadeiros Negativos (VN), Falsos Positivos (FP) e Falsos Negativos (FN). A figura 1 mostra como é feita a distribuição destes valores em um cenário binário sim ou não;

		Detectada	
		Sim	Não
Real	Sim	Verdadeiro Positivo (VP)	Falso Negativo (FN)
	Não	Falso Positivo (FP)	Verdadeiro Negativo (VN)

Fig. 1. Matriz de confusão binária [10].

- Acurácia: Indica um desempenho geral do algoritmo. Essa métrica representa qual foi o percentual de acertos na classificação do algoritmo em relação ao número total de exemplos classificados [10]. A equação 1 apresenta a fórmula de cálculo da acurácia;

$$\text{Acurácia} = \frac{VP + VN}{VP + VN + FP + FN} \quad (1)$$

- Precisão: Dentre todas as classificações de classe Positivo que o modelo fez, indica quantas estão corretas [6]. Um valor alto na precisão indica um baixo número de Falsos Positivos [10]. A equação 2 apresenta a fórmula de cálculo da precisão.

$$\text{Precisão} = \frac{VP}{VP + FP} \quad (2)$$

- Recall/Revocação/Sensibilidade: Dentre todas as situações de classe Positivo como valor esperado, indica quantas estão corretas [6]. Um valor alto de revocação indica uma baixa quantidade de Falsos Negativos [10]. A equação 3 apresenta a fórmula de cálculo da revocação;

$$\text{Revocação} = \frac{VP}{VP + FN} \quad (3)$$

- F-Score: É a média harmônica entre precisão e recall [6]. Um alto valor de F-Score representa uma boa relevância do valor da acurácia, mostrando que os valores de VP, VN, FP, FN não apresentaram grandes distorções. A equação 4 mostra como é representado o F-Score.

$$F_{Score} = \frac{2 * \text{Precisão} * \text{Revocação}}{\text{Precisão} + \text{Revocação}} = \frac{2 * VP}{2 * VP + FP + FN} \quad (4)$$

III. TRABALHOS RELACIONADOS

Existem diversos trabalhos sobre análise de sentimentos utilizando Twitter [22], alguns apresentam a análise de sentimentos usando alguma técnica de visualização [18], [8], [21], outros não [17], [1], [20], [5], [19]. Por exemplo, o trabalho de Khun and Thant [8] apresenta um modelo para análise de sentimento que utiliza uma abordagem léxica junto com uma visualização geográfica, conforme mostra a figura 2. O objetivo era ajudar as pessoas a compreenderem melhor as mudanças de reações de sentimento do público com o caso de

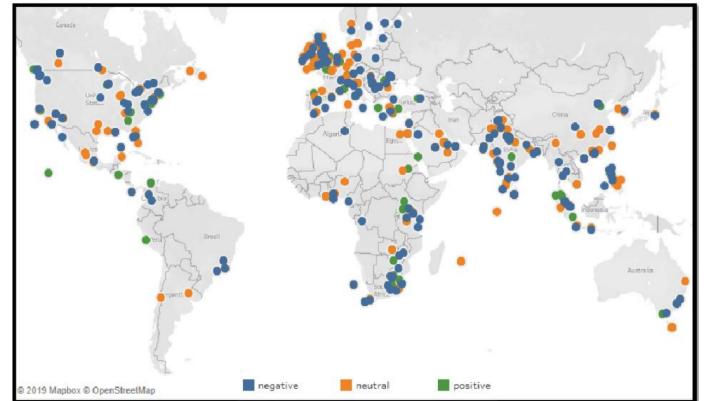


Fig. 2. Visualização do trabalho de Khun and Thant [8], que apresenta um mapa com pontos de diferentes cores. Cada cor representa um sentimento entre negativo, positivo e neutro em cada país. Todos os tweets são relacionados com o caso de banimento da empresa Huawei.

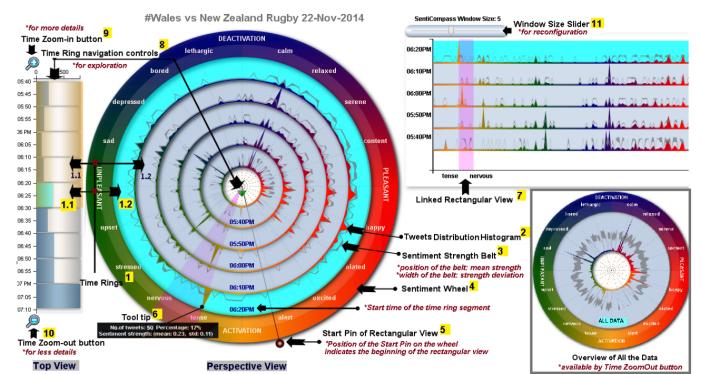


Fig. 3. Visualização proposta por Wang et al. [18], na qual tweets coletados durante uma partida de Rugby do País de Gales x Nova Zelândia foram classificados ao longo do tempo da partida utilizando a ferramenta.

banimento da empresa Huawei dos Estados Unidos por parte do governo americano.

Já o artigo de Wang et al. [18] apresenta um modelo que utiliza uma abordagem híbrida: léxica e aprendizado de máquina utilizando o algoritmo *Naive Bayes*. Na abordagem léxica foi utilizado um dicionário contendo 1034 palavras em inglês e avaliado estatisticamente cada palavra do dicionário contida em um texto analisado. A abordagem utilizando aprendizado de máquina foi utilizada como complemento para compensar a abordagem léxica. Sua visualização combina um modelo de psicologia 2D de afeto chamado de circumplex de Russell com uma representação de túnel do tempo. Pode-se dividir a visualização dos tweets através do túnel tempo, no qual cada divisão é representada por um círculo com seu horário de inicio, ou visualizar todos os tweets por um único círculo. O círculo mais externo representa cada sentimento. Cada sentimento identificado em um texto na visualização é distribuído graficamente em cada sentimento do círculo mais externo conforme ilustra a figura 3.

Wehrmann et al. [19] realizaram um estudo com aprendizagem profunda utilizando Redes Neurais Convolucionais

para classificação de sentimentos e para detecção da língua em tweets. Foram utilizados mais de 128 mil tweets nas línguas portuguesa, alemã, espanhola e inglesa, coletados de um conjunto de dados com tweets já classificados em positivo, negativo ou neutro. Foram utilizadas diferentes abordagens e a que apresentou melhor resultado foi a *Non-negative least squares(NNLS)*, com acurácia de 74.43% na classificação de sentimentos.

No artigo de Aguiar et al. [1] é proposto um modelo que combina diversos algoritmos de aprendizado de máquina, sendo eles, *Naive Bayes*, Regressão Logística, *Support Vector Machine*, *Random Forest* e *Decision tree*, e a partir de uma regra de votação se escolhe qual algoritmo utilizar. Já o artigo de Wan and Gao [17] também apresenta um modelo que combina os algoritmos *Support Vector Machine*, *Random Forest*, Regressão Logística e *Naive Bayes*, atribuindo pesos para cada algoritmo conforme as métricas de cada um. Com este peso cada algoritmo tem uma porcentagem de contribuição para o resultado final.

Crescencio et al. [3] realizaram um estudo de análise de sentimento das opiniões sobre a atração Oktoberfest Blumenau no TripAdvisor. Para classificar as opiniões em positivo ou negativo, eles utilizaram um modelo probabilístico baseado no algoritmo *Naive Bayes* e redes neurais utilizando o algoritmo *Word2Vec*. A solução se baseia na abordagem chamada *Long Short-Term Memory Units – LSTM*, na qual este tipo de rede processa a entrada em sequência, iterando pelos elementos e produzindo uma sequência de saída [3]. O modelo com Redes neurais obteve uma acurácia de 92% e *Naive Bayes* 90%.

Ao analisar os trabalhos relacionados, observamos que a abordagem de aprendizado de máquina é a mais utilizada, sendo o algoritmo *Naive Bayes* o mais usado, seguido, respectivamente, pelos algoritmos *Support Vector Machine*, Regressão Logística e *Random Forest*.

A tabela I apresenta um comparativo entre os trabalhos citados. Dois deles utilizam visualização de dados com abordagens léxicas, sem informação de resultados dos seus modelos de análise de sentimento. Outros dois não disponibilizam técnicas de visualização de dados e utilizam apenas aprendizado de máquina e 2 utilizam Redes Neurais.

IV. METODOLOGIA

Nesta seção apresentamos as etapas realizadas neste trabalho, explicando as hipóteses que gostaríamos de validar e as ferramentas utilizadas.

O trabalho de conclusão de curso realizado por Medina [10] apresenta um estudo sobre classificação de Tweets em português usando aprendizado de máquina. Neste estudo, foi feita uma comparação entre os algoritmos *Support Vector Machine*, Regressão Logística e *Naive Bayes* para classificação de tweets de assuntos diversos e de um assunto específico, no caso, o COVID-19. Para isto um conjunto de dados com cerca de 2.000 tweets foi coletado no primeiro semestre de 2021 e anotado manualmente em positivo, negativo e neutro, e foi dividido pelas seguintes hashtags e quantidades:

- 140 tweets contendo a hashtag **economia**;
- 140 tweets contendo a hashtag **educação**;

Comparativo dos trabalhos relacionados				
Artigo	Algoritmos utilizados	Tweets utilizados	Utiliza visualização	Acurácia
[8]	Abordagem léxica	115.775	Sim	-
[18]	Abordagem léxica, <i>Naive Bayes</i>	122.393	Sim	-
[1]	<i>Naive Bayes</i> , Regressão Logística, <i>Support Vector Machine</i> , <i>Random Forest</i> e <i>Decision tree</i>	2.516	Não	86%
[17]	<i>Support Vector Machine</i> , <i>Random Forest</i> , Regressão Logística e <i>Naive Bayes</i>	12.864	Não	84.2%
[19]	Redes Neurais	128.000	Não	74.43%
[3]	Redes Neurais e <i>Naive Bayes</i>	-	Não	92% e 90%

Tabela I
SÍNTSE DOS TRABALHOS RELACIONADOS.

- 140 tweets contendo a hashtag **entretenimento**;
- 140 tweets contendo a hashtag **saúde**;
- 140 tweets contendo a hashtag **política**;
- 300 tweets contendo a hashtag **netflix**;
- 1000 tweets contendo a hashtag **covid19**.

Neste trabalho, damos continuidade a este estudo [10], aumentando o conjunto de dados e fazendo uma combinação dos algoritmos de classificação utilizados com o objetivo de obter resultados melhores do que com os algoritmos individuais como visto no artigo de Aguiar et al.[1] . As três hipóteses principais que norteiam nosso trabalho são:

- 1): Somente com o aumento do conjunto de dados os três algoritmos (*Support Vector Machine*, Regressão Logística e *Naive Bayes*) terão resultados melhores;
- 2): Através da combinação dos três algoritmos (*Support Vector Machine*, Regressão Logística e *Naive Bayes*) será possível classificar tweets em português de assuntos diversos de forma satisfatória, com uma acurácia maior ou igual a 70%;
- 3): Através da combinação dos três algoritmos (*Support Vector Machine*, Regressão Logística e *Naive Bayes*) os resultados serão melhores que os três algoritmos individualmente.

Para validação, ou não, dessas hipóteses, dividimos nosso trabalho em 7 etapas:

- 1) Coleta de tweets;
- 2) Pré-processamento e anotação;
- 3) Junção e duplicação dos conjuntos de dados, através da qual o conjunto de dados do estudo de Medina [10] foi combinado com nosso conjunto de dados, havendo, então, uma duplicação do conjunto de dados;
- 4) Treinamento dos algoritmos *Support Vector Machine*, Regressão Logística e *Naive Bayes* e comparação com os resultados obtidos no estudo de Medina [10] ;
- 5) Tratamento de desbalanceamento dos dados;

- 6) Criação do nosso modelo de aprendizado de máquina combinando os três algoritmos;
- 7) validação.

Outras 2 etapas completam a metodologia do nosso trabalho sendo elas a implementação da visualização de dados e integração do nosso modelo de aprendizado de máquina com a ferramenta PeakVis [14].

A figura 4 tem como objetivo ilustrar todas as etapas do nosso trabalho.

O desenvolvimento das rotinas de coleta de dados, pré-processamento de tweets e treinamento/teste pelos algoritmos foi realizado na plataforma Jupyter Notebook, utilizando a linguagem de programação Python. Diversas bibliotecas feitas para Python foram utilizadas no trabalho e as principais são: Tweepy², para acessar a API do Twitter e realizar a coleta de tweets; NLTK³, para as funções de pré-processamento; e scikit-learn⁴, para as funções de aprendizado de máquina.

V. CONJUNTO DE DADOS

Nesta seção apresentamos o conjunto de dados utilizado no nosso estudo, com o objetivo de ampliar o entendimento de como os dados estão distribuídos e explicar em maiores detalhes as etapas de coleta de dados, pré-processamento e anotação e, por fim, junção e duplicação dos conjuntos de dados.

A. Coleta de dados

A coleta de dados foi realizada usando a API pública do Twitter, através do desenvolvimento de scripts. Nesta etapa utilizamos a biblioteca Tweepy, que permite o acesso a API do Twitter. Com o auxílio desta biblioteca, basta fornecer alguns parâmetros a API, tais como: as palavras de busca, o número de tweets que desejamos coletar e a língua presente nos tweets[10]. No nosso trabalho foi utilizada a língua portuguesa e em todos os tweets foram filtrados os retuítes, para não haver dados repetidos assim como no estudo de Medina [10]. Desta maneira, ao total foram coletados 1.700 tweets, todos do segundo semestre de 2021. A seguir apresentamos como eles foram divididos no nosso conjuntos de dados.

- 160 tweets contendo a hashtag **economia**;
- 160 tweets contendo a hashtag **educação**;
- 160 tweets contendo a hashtag **entretenimento**;
- 160 tweets contendo a hashtag **saúde**;
- 160 tweets contendo a hashtag **política**;
- 300 tweets contendo a hashtag **bbb20, masterchef**;
- 300 tweets contendo a hashtag **amordemae ou adonadopedaco** ambas nome de novelas;
- 300 tweets contendo a hashtag **olimpiadas21 ou grenal**;

Assim ao realizar a combinação dos dados com o conjunto de dados do estudo de Medina [10], aumentamos de 140 tweets nas hashtags **economia, educação, entretenimento, saúde e política** para 300 tweets. Além disso, adicionamos novas hashtags: 300 de uma categoria que chamamos de *reality*

²<https://www.tweepy.org>

³<https://www.nltk.org>

⁴<https://scikit-learn.org/stable>

show, dividindo 150 tweets para o programa Big Brother Brasil e 150 tweets para o programa MasterChef Brasil; 300 tweets para uma categoria chamada *novela*, dividindo 150 para a novela “Amor de Mãe” e 150 para a novela “A Dona do Pedaço”; e, por fim, 300 tweets para a categoria esporte, dividindo 150 para as olimpíadas de 2021 e 150 para um jogo de futebol. No final, totalizamos 3.700 tweets.

B. Pré-processamento e anotação

Após a coleta dos tweets foi preciso fazer a anotação manual. Para isto, consideramos as três categorias de sentimentos já utilizadas no conjunto de dados do estudo de Medina [10]: positivo, negativo e neutro. Cada tweet é classificado apenas em uma dessas categorias. Neste caso, torna-se importante entender como é o processo de anotação desses tweets.

No estudo de Medina [10] o processo de anotação dos dados foi feito de forma manual, sendo realizado por três pessoas sem experiência em anotação de tweets. Neste caso, duas pessoas classificaram todos os tweets seguindo determinados critérios para cada classe e caso houvesse algum tweet com classificação diferente a terceira pessoa realizava a classificação.

Nosso trabalho seguiu a mesma metodologia do estudo de Medina [10], se diferenciando nas pessoas que realizaram a anotação dos tweets que não foram as mesmas, mas que também não tinham experiência em anotação de tweets, e adicionando novas hashtags e critério de classificação para as mesmas. As hashtags **economia, educação, entretenimento, saúde e política** seguiram os mesmos critérios apresentados em Medina [10]. A tabela II exemplifica estes critérios.

Com o objetivo de verificarmos a qualidade de anotação dos dados, realizamos o processo de analisar a quantidade de tweets que um terceiro anotador teve que realizar. A tabela V-B exibe por hashtag a porcentagem de tweets em que houve a necessidade de um terceiro anotador. Como podemos observar, em assuntos relacionados à saúde, economia e política houve uma maior quantidade de tweets em que foi necessário um terceiro anotador. No geral, observamos um baixo número de tweets tendo que ser anotado mais de 2 vezes o que nos leva a concluir que os critérios adotados ajudaram na qualidade do processo de anotação.

A etapa seguinte é de pré-processamento dos tweets. Essa etapa consiste na aplicação de uma série de técnicas com o objetivo de excluir informações irrelevantes e deixar o texto mais fácil de ser analisado pelos algoritmos [5]. Nesta etapa utilizados a biblioteca NLTK. Foram utilizadas as seguintes técnicas de pré-processamento textual:

- **Remoção de stop Words:** Exclusão de palavras que não agregam sentido à análise, como artigos e preposições. Por exemplo, a frase “Eu gosto muito de ir ao Shopping e sair com os meus amigos.”, após a remoção das stop Words ficaria: “gosto muito ir Shopping sair amigos”.
- **Stemização:** Redução de palavras a sua raiz, diminuindo o número total de palavras diferentes. Assim, por exemplo, palavras como gostava, gostei, gostando, gosta serão reduzidas a apenas “gost”.
- **Tokenização:** Separação de cada termo nos tweets em unidades mínimas, sem perda do sentido original do

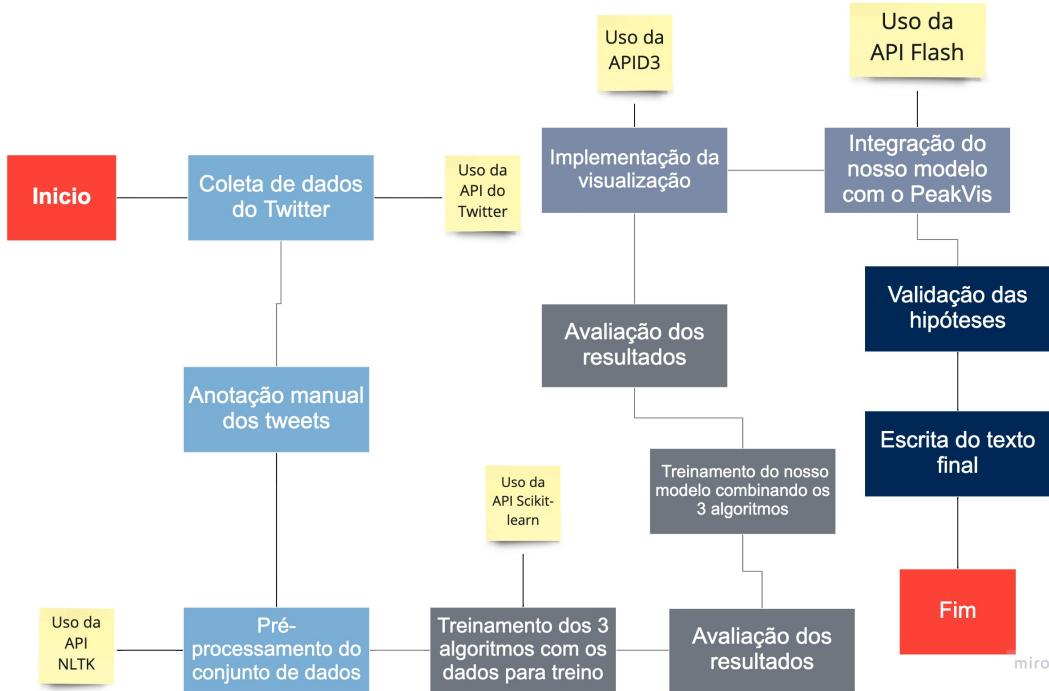


Fig. 4. Etapas do desenvolvimento do trabalho.

tweet [10]. Por exemplo: Eu amo viajar para a praia. Seria gerada uma lista contendo 6 tokens: 'Eu', 'amo', 'viajar', 'para', 'a', 'praia'. A biblioteca NLTK apresenta uma função específica de tweets chamada *TweetTokenizer*, que permite uma tokenização mais eficiente, pois leva em conta termos específicos que aparecem no Twitter como caracteres '@', emoji e hashtag.

- **Tratamento de caracteres:** Exclusão de símbolos, acentos, pontuações, números e links. Para isto foram utilizadas expressões regulares.

A aplicação das técnicas Remoção de stop Words, Stemização e Tokenização, aumentam a qualidade da análise textual [11] e melhoram a questão da falta de formalismo de texto online [2].

Nosso trabalho utiliza a análise dos tweets a nível de sentença, na qual cada frase é analisada separadamente e são classificadas conforme a sua polaridade positiva, negativa ou neutra [3].

C. Junção e duplicação dos conjuntos de dados

Esta etapa é importante para validação das hipóteses, através da qual geramos três conjuntos de dados, sendo o primeiro a combinação dos 2.000 tweets do estudo de Medina [10] com os 1.700 tweets coletados no nosso trabalho. O segundo conjunto de dados consiste na combinação dos 2.000 tweets do estudo de Medina [10] com 800 tweets do nosso trabalho sendo eles os tweets das hashtags: **economia, educação, entretenimento, saúde e política**. Estes dois primeiros conjuntos de dados foram utilizados para verificar a validação da hipótese (1). No primeiro conjunto de dados temos, além do aumento de dados das hashtags **economia, educação, entretenimento,**

saúde e política, novas hashtags, como *bbb20* e *masterchef*. Já no segundo conjunto de dados temos apenas o aumento de dados nas hashtags **economia, educação, entretenimento, saúde e política**, sem adição de novas hashtags.

No terceiro conjunto de dados, que auxilia na validação das hipóteses (2) e (3), combinamos 600 tweets do estudo de Medina [10], sendo 300 tweets da hashtag *netflix* e 300 de *covid19*, com 1.700 tweets coletados no nosso trabalho. Desta forma, não tendo tweets da mesma hashtag entre o conjunto de dados do estudo de Medina [10] e o que coletamos no nosso trabalho. A tabela IV ilustra os três conjuntos de dados e suas respectivas distribuições em cada categoria de sentimento.

D. Análise do tweets

Para nos ajudar a entender como cada hashtag influencia em cada categoria de sentimento, primeiramente separamos o conjunto de dados 1 em cada hashtag e conforme podemos observar na figura 5 que representa todos os tweets negativos, neutros e positivos para cada hashtag. Tendo Covid-19, Política, Netflix e Educação com a maior representação no negativo. Olimpiadas, Economia, Covid-19 e Saúde possuindo maior representação no positivo. E por fim, Entretenimento, Educação, Covid-19 e Saúde possuindo maior representação no neutro.

Após este processo de separação do conjunto de dados por hashtags, buscamos entender quais as principais ideias que estavam presentes nos conjuntos de dados. Para isto criamos nuvens de palavras com os termos presentes em cada conjunto de dados. A criação de nuvens de palavras permite visualizar quais palavras aparecem com maior frequência em um texto, de forma a dar uma visão melhor sobre quais termos e ideias

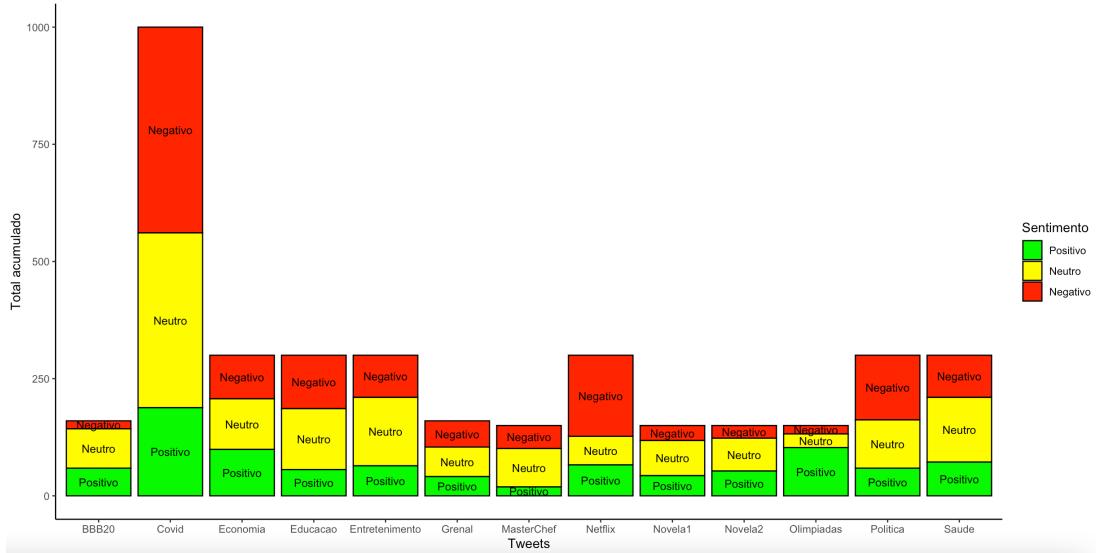


Fig. 5. Gráfico distribuição das *Hashtags* por sentimento.

estão presentes no conjunto de dados [10]. Além disso, a construção de nuvens de palavras permite uma análise dos dados mais detalhada e segura [16]. A tabela V ilustra quais as palavras mais frequentes em cada um dos conjuntos de dados por sentimento. E podemos observar que no conjunto de dados 1 temos a palavra vacina com destaque no positivo e seu surgimento também com certo destaque no negativo. Já no conjunto de dados 2 temos a palavra dose com destaque no positivo o que acreditamos se referir a aplicação da vacina do covid-19 e a palavra pandemia com destaque no negativo. E por fim, no conjunto de dados 3 temos a palavra ouro com destaque no positivo no qual acreditamos se referir a medalha de ouro nas olimpíadas e a palavra covid com destaque no negativo.

VI. MODELO DE ANÁLISE DE SENTIMENTOS

Nesta seção apresentamos o processo de treinamento dos algoritmos de aprendizado de máquina e nosso modelo desenvolvido que combina três algoritmos de aprendizado de máquina.

A. Treinamento dos algoritmos

Primeiramente, para que os algoritmos *Support Vector Machine*, Regressão Logística e *Naive Bayes* possam ser treinados e testados, é necessário que os dados passem por uma etapa chamada extração de *features*. Essa etapa é importante pois ela é responsável por transformar o texto dos tweets em algo que os algoritmos possam entender [10]. No estudo de Medina [10] foi utilizado a abordagem de *Bag of Words*. No qual, seu funcionamento consiste em inicialmente gerar uma lista de tamanho n com todos os termos do conjunto de dados. Após isso, cada um dos tweets é transformado em um vetor esparsa com n posições, e cada posição deste vetor representa o número de vezes que cada termo das n posições aparece no tweet. Utilizamos a mesma abordagem do trabalho

de Medina [10] e a biblioteca *CountVectorizer*⁵, pelo fato dela apresentar uma maneira simples de realizar a extração de *features* utilizando a abordagem *Bag of Words* e também ter sido utilizada no trabalho de Medina [10]

Após a realização da etapa de extração de *features*, os três algoritmos já estão prontos para serem treinados e testados. Para treinamento e teste utilizamos a técnica *Hold-Out*. *Hold-Out* é uma maneira de se fazer a validação cruzada, isto é, verificar o poder de generalização do modelo [10]. Nesta técnica, o conjunto de dados é dividido em conjunto de treinamento, normalmente com 70% dos dados e 30% para testes [10].

B. Tratamento de desbalanceamento dos dados

Dados Desbalanceados podem ser definidos pela maior quantidade de uma determinada categoria dentro de um conjunto de dados em comparação com as demais categorias [9]. Isto faz com que tenhamos muitas informações a respeito das categorias com maior quantidade, e menos das demais, o que pode, em muitos casos, interferir nos resultados dos algoritmos de aprendizado de máquina [9]. Uma forma de tirar o viés causado pela diferença de proporção das categorias consiste em manipular a quantidade de dados que são efetivamente utilizados pelo modelo de aprendizado de máquina, tentando igualar o número de observações entre as categorias [9]. Duas abordagens conhecidas são:

- *Undersampling*: Esse método consiste em reduzir o número de observações da classe majoritária para diminuir a diferença entre as categorias [9].
- *Oversampling*: Consiste em criar sinteticamente novas observações da classe minoritária, com o objetivo de igualar a proporção das categorias [9].

Considerando nossa primeira hipótese, apresentada na Seção IV, na qual compararmos resultados com o estudo de Med-

⁵https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html

Critério de anotação dos tweets			
Categoria	Positivo	Negativo	Neutro
Economia	Tweets que apresentam, melhorias nos indicadores da economia.	Tweets que apresentam, indicativos de piora ou críticas à economia.	Tweets que não apresentam polaridade.
Saúde	Tweets que apresentam, Melhorias nos indicadores de saúde.	Tweets que apresentam, informação sobre piora ou critica a saúde.	Tweets que não apresentam polaridade.
Educação	Tweets que apresentam, Melhorias nos indicadores de Educação.	Tweets que apresentam, informação sobre piora ou crítica a Educação.	Tweets que não apresentam polaridade.
Entretenimento	Tweets que apresentam, elogio a algum evento ou celebridade.	Tweets que apresentam, crítica a algum evento ou celebridade.	Tweets que não apresentam polaridade ou apenas uma informação imparcial.
Política	Tweets que apresentam, elogios a algum político ou partido ou alguma ação feita pelo político ou partido.	Tweets que apresentam notícias sobre escândalos na política ou crítica de algum político ou partido.	Tweets que não apresentam polaridade como o caso de notícias imparciais.
Covid-19	Tweets que apresentam, informação sobre a baixa de casos de covid-19 ou número de vítimas, tweets sobre a vacinação e com tons de esperança.	Tweets que apresentam, informação sobre a alta de casos de covid-19 ou número de vítimas, tweets sobre a falta de cuidado ou descaso sobre covid.	Tweets que não apresentam polaridade como o caso de notícias como as de ações para melhorar a saúde do estado do RS.
Netflix	Tweets que apresentam, elogio a algum filme ou série ou a própria netflix.	Tweets que apresentam, críticas a personagens, séries ou filmes ou a própria netflix.	Tweets que não apresentam polaridade como o caso de notícias como "Lançamentos da semana netflix".
Reality show	Tweets que apresentam, elogio a algum participante ou ao programa em si.	Tweets que apresentam, críticas ao programa ou a algum participante.	Tweets que não apresentam polaridade como o caso de notícias.
Novelas	Tweets que apresentam, elogio a algum personagem ou a novela em si.	Tweets que apresentam, críticas a algum personagem ou a novela em si.	Tweets que não apresentam polaridade.
Esportes	Tweets que apresentam elogios a algum atleta ou esporte e comemorações de vitórias.	Tweets que apresentam crítica a algum atleta ou esporte.	Tweets que não apresentam polaridade.

Tabela II
CRITÉRIO DE ANOTAÇÃO DOS TWEETS.

Categoría	Total de tweeets(n)	Total de tweeets(%)
economia	43	26%
educação	11	6.8%
entretenimento	14	8.7%
saúde	48	30%
politica	51	31,8%
bbb20	9	6%
masterchef	15	10%
amordemae	27	18%
adonadopedaco	19	12.6%
olimpiadas21	17	11.3%
grenal	21	14%

Tabela III
TOTAL DE TWEETS ANOTADOS POR UM TERCEIRO ANOTADOR POR CATEGORIA.

Sentimento	Conjunto de dados 1	Conjunto de dados 2	Conjunto de dados 3
Positivo	25%	21.42%	27.21%
Negativo	35.46%	40.42%	26.96%
Neutro	39.54%	38.16%	45.83%
Total	3700	2800	2300

Tabela IV
SÍNTSE DOS CONJUNTOS DE DADOS.

ina [10], cujo em seu trabalho não foi realizado nenhum tratamento de desbalanceamento dos dados, para validação ou não das hipóteses (2) e (3), decidimos inicialmente realizar o balanceamento dos dados afim de não termos uma categoria influenciando mais que outras o treinamento dos algoritmos. Optamos pelo método *Undersampling* pela sua simplicidade de implementação no qual removemos dados das categorias com maiores quantidades de forma aleatória, mesmo que com isso poderíamos perder tweets que ajudariam no treinamento dos algoritmos. Como observamos na tabela IV que apresenta as quantidades de tweets em cada conjunto de dados, em nosso conjunto de dados 3 a maioria dos dados se concentram na classe Neutra. Por isto, geramos um novo conjunto de dados que chamamos de conjunto de dados 4, o qual contém 620 tweets de cada categoria.

C. Criação e treinamento do nosso modelo de aprendizado de máquina

Segundo Faceli [6] podemos combinar algoritmos de aprendizado com o objetivo de obtermos melhores resultados que os algoritmos individuais. Este processo é chamado de *Ensemble Learning*, ou chamado de aprendizado por agrupamento, e se baseia na ideia de combinar diversos modelos de predição mais simples, treiná-los para uma mesma tarefa, e produzir a partir desses um modelo agrupado mais complexo que é a soma de suas partes Faceli [6]. Existem diversas técnicas de combinação de algoritmos Faceli [6], como o algoritmo *Bagging* que utiliza um algoritmo de aprendizado de máquina como Regressão logística ou *Support Vector Machine*, divide o conjunto de dados em N partes e para cada uma destas parte cria um modelo de aprendizado de máquina utilizando o algoritmo escolhido Faceli [6]. Após o treinamento de todos os modelos, a cada nova classificação é realizada uma média



Tabela V
NUVEM DE PALAVRAS DOS CONJUNTOS DE DADOS.

somando todos os modelos Faceli [6]. Esta técnica tenta evitar termos um modelo com algum viés nos dados ao termos todo o conjunto de dados treinado por modelos diferentes mesmo que utilizando o mesmo algoritmo Faceli [6].

Outra técnica vista no trabalho de Aguiar et al. [1], consiste em treinar algoritmos de aprendizado de máquina de forma individual e agrupar estes algoritmos a cada nova classificação, realizando uma votação seguindo algum critério definido. Em seu trabalho o mesmo utilizou como critério a acurácia de cada modelo individual, assim alguns algoritmos tinham maior peso na votação do que outros.

Já o estudo de Medina [10] demonstra que algoritmos de aprendizado de máquina obtêm melhores resultados quando treinados para resolver problemas binários. Neste trabalho, foi utilizado um conjunto de dados de tweets sobre Covid-19 para classificar sentimentos em positivo, negativo ou neutro utilizando algoritmos de aprendizado de máquina *Support Vector Machine*, Regressão Logística e *Naive Bayes*. E foi demonstrado que ao treinar os algoritmos apenas para classificar entre positivo e negativo se obteve uma acurácia de 81.13%, enquanto utilizando os três sentimentos se obteve uma acurácia de 62.17%.

Desta forma propomos um modelo de aprendizado de máquina que combina os algoritmos *Support Vector Machine*, Regressão Logística e *Naive Bayes* e utiliza a técnica de votação e treinamento binário para classificar sentimentos em tweets. Este classificador será *multiclass* ou seja poderá classificar em mais de um sentimento e *single label* apenas 1 sentimento será escolhido. Para isto, inicialmente, após termos o conjunto de dados balanceado, nós dividimos os dados da seguinte maneira: POSITIVO X NEGATIVO, POSITIVO

X NEUTRO e NEGATIVO X NEUTRO transformando em conjunto de dados binário. Após esta etapa treinamos os três algoritmos utilizando a técnica *Hold-out*, em cada um dos conjuntos de dados. A tabela VI-C exibe os resultados de cada algoritmo em cada conjunto de dados nas métricas de precisão, acurácia, F-Score e *recall*. Com esta etapa concluída escolhemos três algoritmos para compor o nosso modelo seguindo os seguintes critérios:

- Nosso modelo pode ter apenas um algoritmo de cada conjunto de dados, sendo este algoritmo diferente entre os conjuntos de dados, ou seja, se Regressão logística for escolhido pelo conjunto de dados POSITIVO X NEGATIVO o algoritmo escolhido pelos conjuntos de dados POSITIVO X NEUTRO e NEGATIVO X NEUTRO não poderão ser Regressão logística. Desta forma, tentamos evitar qualquer viés causado pela escolha do mesmo algoritmo especialmente nos conjuntos de dados POSITIVO X NEUTRO e NEGATIVO X NEUTRO por ambos terem os mesmos dados no sentimento Neutro.
- O algoritmo de cada conjunto de dados será escolhido baseado no que apresentar melhor F-Score. A escolha do F-Score se deu por esta métrica informa o quanto preciso é o classificador, ou seja, quantos dados ele classifica corretamente, bem como o quanto robusto ele é ou seja seu *Recall*. Um bom valor de F-Score acaba mostrando que os valores de VP, VN, FP e FN não apresentaram grandes distorções Medina [10].
- Caso o algoritmo já tenha sido escolhido por um conjunto de dados o segundo melhor colocado no F-Score que não tenha sido escolhido por outro conjunto de dados é utilizado e assim sucessivamente.

Conjunto de dados	Algoritmo	Precisão	Recall	F-Score	Acurácia
POSITIVO X NEGATIVO	Support Vector Machine	75%	77%	76%	76%
	Naive Bayes	74%	72%	73%	73%
	Regressão Logística	76%	75%	74%	75%
POSITIVO X NEUTRO	Support Vector Machine	69%	67%	68%	68%
	Naive Bayes	72%	62%	67%	70%
	Regressão Logística	67%	70%	69%	69%
NEGATIVO X NEUTRO	Support Vector Machine	69%	67%	68%	68%
	Naive Bayes	71%	70%	70%	70%
	Regressão Logística	70%	71%	69%	70%

Tabela VI
COMPARATIVOS COM CONJUNTO DE DADOS BINÁRIOS.

Algorithm 3: Modelo proposto

```

Data: tweet
Result: Classificação
classificacao1 = POSITIVO X NEGATIVO Support
Vector Machine(tweet);
classificacao2 = POSITIVO X NEUTRO Regressão
Logística(tweet);
classificacao3 = NEGATIVO X NEUTRO Naive
Bayes(tweet);
if classificacao2 == classificacao3 then
| return classificacao2;
else
| if classificacao2 == classificacao1 then
| | return classificacao2;
| else
| | if classificacao3 == classificacao1 then
| | | return classificacao3;
| | else
| | | return classificacao1;
| end
end

```

end

Fig. 6. Pseudo-algoritmo desenvolvido.

Para treinar nosso modelo, utilizamos 30% dos tweets de cada conjunto de dados binário, tendo um total de 538 tweets.

VII. VISUALIZAÇÃO DOS TWEETS

Nesta seção apresentamos a visualização de dados desenvolvida e a integração do nosso trabalho com a ferramenta PeakVis.

A. Implementação da visualização de dados

A visualização da classificação dos tweets foi desenvolvida utilizando a linguagem de programação Javascript e a biblioteca D3⁶ que fornece diversas visualizações para uso de forma mais facilitada.

Para definição de qual visualização seria a mais adequada para usar, inicialmente levantamos algumas características que a visualização deveria ter considerando os dados, ou seja, tweets classificados em positivo, neutro e negativo. Assim, chegamos nas seguintes características:

- Fácil distinção entre as categorias de sentimentos;
- Fácil identificação de tweets relevantes baseados nos seus retweets;

Com essas características nós optamos pela visualização Beeswarm⁷, a qual consiste em um gráfico unidimensional, em outras palavras, um gráfico que mostra todas as informações em um único eixo. Ele exibe valores como uma coleção de pontos. Esse tipo de gráfico é muito útil quando se deseja exibir muitos pontos de dados de uma só vez.

⁶<https://d3js.org/>

⁷<https://flourish.studio/2019/12/03/beeswarm-violin-plot-webgl-scatter/>

Conforme os critérios descritos, os algoritmos escolhidos foram *Support Vector Machine* para POSITIVO X NEGATIVO, Regressão Logística para POSITIVO X NEUTRO e *Naive Bayes* para NEGATIVO X NEUTRO. Com os algoritmos escolhidos combinamos os 3 algoritmos e realizamos a técnica de votação, na qual a cada tweet informado para se classificar, cada algoritmo informa sua classificação e o sentimento mais votado é o escolhido a figura 6 exibe o pseudo-algoritmo desenvolvido. Desta forma, utilizamos a vantagem de termos modelos com bons resultados binários para classificar de forma não binária.

A Figura 7 ilustra a visualização desenvolvida. Cada círculo/ponto representa um tweet, sendo o seu tamanho definido pela quantidade de reweets. Para isto, ao iniciar a visualização com um conjunto de dados é buscado qual é o valor do maior reweet e, então, este valor é dividido em 5 partes definindo 5 tamanhos. Já cada cor define um sentimento sendo: negativo na cor vermelha, positivo na cor verde e neutro na cor amarela. A posição de cada categoria de sentimento na visualização é definida de forma fixa no eixo Y, no qual utilizando funções da biblioteca D3 para bloquear a colisão dos elementos no gráfico. Nossa visualização também permite filtrar os dados por quantidade de reweets, por sentimento ou por palavras.

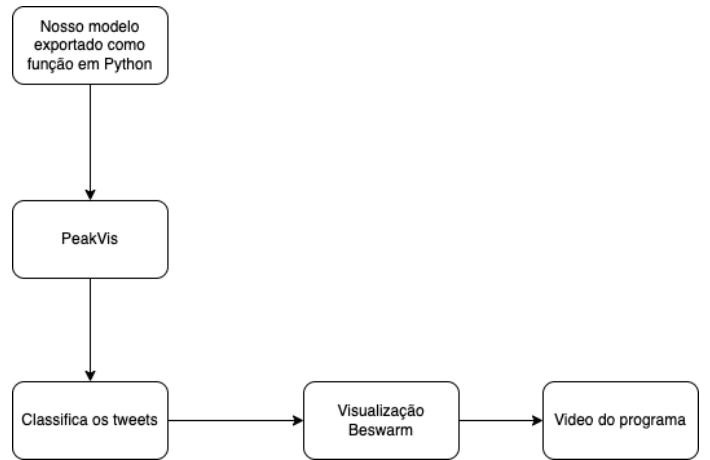


Fig. 8. Integração com o PeakVis.

B. Integração com a ferramenta PeakVis

Uma das características do PeakVis é permitir a análise de tweets baseado em programas *broadcast*. Levando esta característica em consideração, nossa integração com o PeakVis sincroniza os dados da nossa visualização com a gravação do vídeo do programa *broadcast*. Para isto, primeiramente todos os dados são ordenados pelo seu horário e em seguida através de funções disponibilizadas pela linguagem Javascript para manipulação de vídeo, a cada x segundos novos tweets relacionados aquele período de tempo são exibidos. A definição de em quantos segundos serão carregados novos tweets está diretamente relacionada ao tempo total do vídeo(TTV) e à quantidade de tweets(QT). A equação 1 exibe como é definido em quantos segundos deve-se exibir novos tweets.

$$\text{segundos} = QT/TTV \quad (1)$$

Já a integração do nosso modelo de aprendizado de máquina com o PeakVis foi feita através da criação de uma função que internamente contém nosso modelo de aprendizado de máquina, que ao informar um texto a mesma retorna a classificação desse texto. O processo de integração foi relativamente simples. Logo após exportar nosso modelo através de uma função para o PeakVis, utilizamos a biblioteca D3 em uma área da página HTML do PeakVis na qual foi renderizada nossa visualização. Antes da renderização da visualização o PeakVis classifica os tweets em positivo, negativo e neutro e informa essa classificação para a visualização.

VIII. RESULTADOS

Nesta seção apresentamos os resultados do nosso trabalho baseado nas hipóteses que gostaríamos de validar, dois estudos de casos utilizando nosso modelo de aprendizado de máquina e o uso do modelo com o PeakVis.

A. Validação das Hipóteses

Para validar ou não a primeira hipótese utilizamos o conjunto de dados 1 e 2 (IV) e comparamos com os resultados obtidos no trabalho de Medina [10]. Para isto, o experimento realizado foi mesmo apresentado no trabalho de Medina [10]: fazer o treinamento com um conjunto de dados de assuntos diversos e testar em um conjunto de dados com uma *hashtag* específica não usada no treinamento, no caso Netflix e covid-19. A tabela VIII-A exibe os resultados obtidos, após o treinamento e testes dos algoritmos na *Hashtag* Netflix e a tabela VIII-A apresenta os resultados obtidos na *Hashtag* Covid-19.

Afim de entendermos se o aumento de novas *hashtags* ou seja novos assuntos, pode influenciar no treinamento dos algoritmos para validação da primeira hipótese e se assuntos como economia, saúde, e política que tem tweets do primeiro e segundo semestre de 2021, também podem influenciar pela mudança de palavras e opiniões sobre determinado assunto como COVID-19. Realizamos um novo experimento, no qual utilizamos a mesma abordagem para validar a hipótese (1) alterando o conjunto de dados.

Dois novos conjuntos de dados foram gerados utilizando tweets dos conjuntos de dados apresentados na Seção V(IV) um que chamamos de Trabalho Medina + novas *hashtags* no qual combinados os tweets do estudo de Medina [10] com os tweets coletados de novas *hashtags*. E o outro composto apenas de dados coletados no segundo semestre de 2021 nas *hashtags* economia, política, saúde, entretenimento e educação que chamamos de tweets segundos semestre. A tabela VIII-A exibe os resultados obtidos, após o treinamento e testes dos algoritmos na *Hashtag* Netflix e a tabela VIII-A apresenta os resultados obtidos na *Hashtag* Covid-19.

TWEET EMOTION CHART

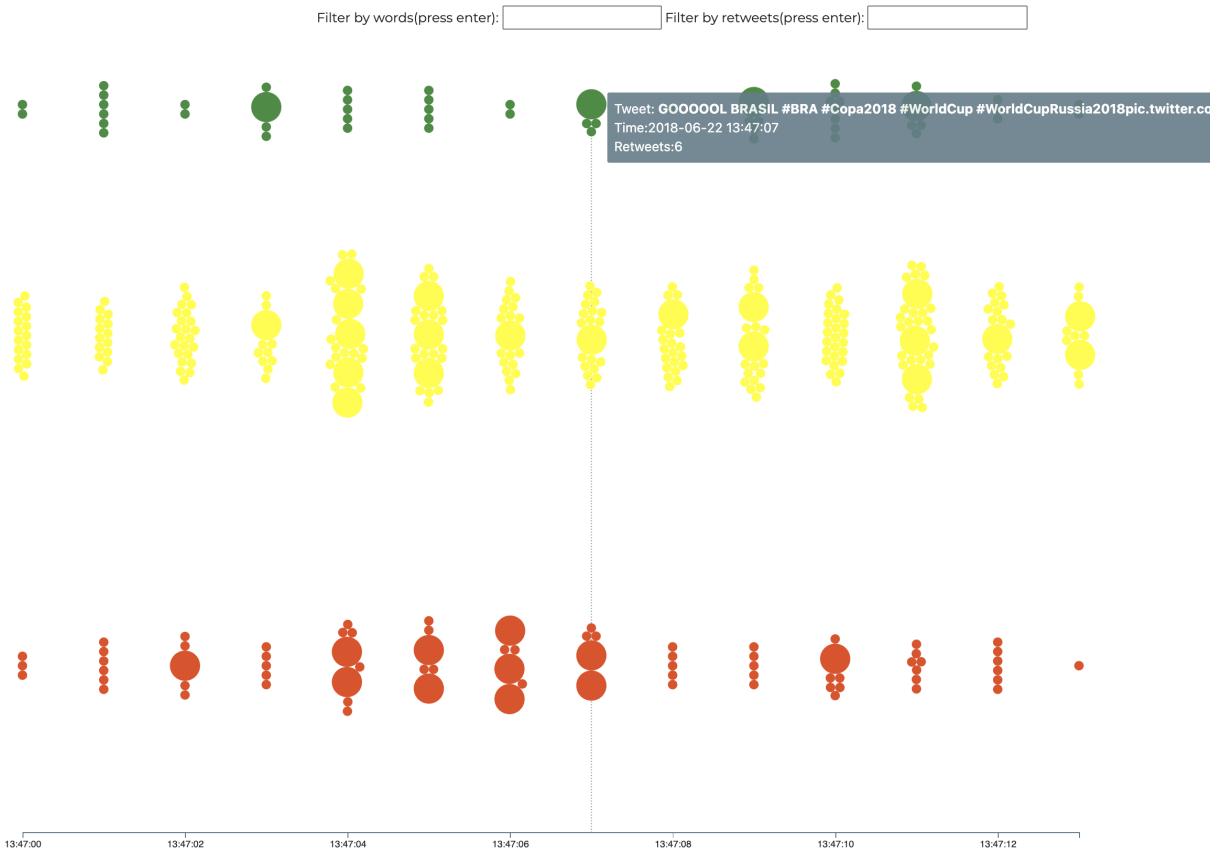


Fig. 7. Visualização de sentimentos.

Conjunto de dados	Algoritmo	Precisão	Recall	F-Score	Acurácia	Conjunto de dados	Algoritmo	Precisão	Recall	F-Score	Acurácia
Medina [10]	Support Vector Machine	54%	54%	52%	54%	Medina [10]	Support Vector Machine	53%	57%	50%	56.67%
	Naive Bayes	56%	56%	55%	56%		Naive Bayes	51%	56%	51%	55.67%
	Regressão Logística	56%	56%	56%	54%		Regressão Logística	57%	59%	49%	58.67%
Conjunto de dados 1	Support Vector Machine	48%	49%	48%	46%	Conjunto de dados 1	Support Vector Machine	29%	28%	25%	28%
	Naive Bayes	45%	49%	47%	46%		Naive Bayes	35%	30%	26%	27%
	Regressão Logística	50%	52%	50%	48%		Regressão Logística	32%	31%	28%	30%
Conjunto de dados 2	Support Vector Machine	47%	49%	47%	47%	Conjunto de dados 2	Support Vector Machine	36%	36%	35%	38%
	Naive Bayes	50%	50%	50%	52%		Naive Bayes	42%	43%	42%	47%
	Regressão Logística	50%	52%	50%	50%		Regressão Logística	41%	42%	40%	44%

Tabela VII

COMPARATIVOS CONJUNTO DE DADOS COM Hashtag COVID.

Tabela VIII

COMPARATIVOS CONJUNTO DE DADOS COM Hashtag NETFLIX.

Conjunto de dados	Algoritmo	Precisão	Recall	F-Score	Acurácia	Algoritmo	Sentimento	Precisão	Recall	F Score	Acurácia
Medina + novas hashtags	Support Vector Machine	47%	49%	47%	47%	Support Vector Machine	Positivo	60%	60%	60%	58%
	Naive Bayes	50%	50%	50%	52%		Negativo	59%	60%	59%	58%
	Regressão Logística	50%	52%	50%	50%		Neutro	54%	53%	54%	58%
tweets segundos semestre	Support Vector Machine	39%	39%	36%	35%	Naive Bayes	Positivo	58%	66%	62%	58%
	Naive Bayes	37%	42%	38%	35%		Negativo	58%	63%	60%	58%
	Regressão Logística	42%	42%	39%	37%		Neutro	60%	46%	52%	58%
No nosso Algoritmo	Positivo	60%	64%	62%	52%	Regressão Logística	Positivo	60%	64%	62%	52%
	Negativo	61%	61%	61%	52%		Negativo	61%	61%	61%	52%
	Neutro	58%	55%	56%	52%		Neutro	58%	55%	56%	52%
No nosso Algoritmo	Positivo	67%	77%	72%	71%	Nosso Algoritmo	Negativo	75%	72%	73%	71%
	Negativo	75%	72%	73%	71%		Neutro	71%	63%	67%	71%
	Neutro	71%	63%	67%	71%						

Tabela IX

COMPARATIVOS MEDINA + NOVAS hashtags X TWEETS SEGUNDOS SEMESTRE COM Hashtag COVID.

Tabela XI
RESULTADOS DOS ALGORITMOS.

Conjunto de dados	Algoritmo	Precisão	Recall	F-Score	Acurácia
Medina + novas hashtags	Support Vector Machine	36%	36%	35%	38%
	Naive Bayes	42%	43%	42%	47%
	Regressão Logística	41%	42%	40%	44%
tweets segundos semestre	Support Vector Machine	25%	31%	15%	19%
	Naive Bayes	28%	30%	17%	20%
	Regressão Logística	30%	34%	15%	21%

Tabela X

COMPARATIVOS MEDINA + NOVAS hashtags X TWEETS SEGUNDOS SEMESTRE COM Hashtag NETFLIX.

Já para validar ou não a segunda e terceira hipótese, utilizamos o conjunto de dados 3 com os dados balanceados e realizamos o processo de treinamento e teste com a técnica *Hold out* nos três algoritmos e em nosso modelo. A tabela **VIII-A** exibe a comparação entre os algoritmos individuais e o nosso modelo em cada sentimento e nas métricas: precisão, recall, F-score e acurácia.

B. Estudo de caso

Para nosso estudo de caso coletamos tweets referentes a dois assuntos. O primeiro está relacionado às reações ao trailer oficial lançado no dia 16 de Novembro de 2021 do filme “Homem - Aranha sem volta para casa”, considerado

um dos filmes mais aguardados do ano de 2021 segundo o site Adorocinema⁸. Para está coleta utilizamos as hashtags: *homemaranhasemvoltaparacasa* e *spidermannowayhome*. O objetivo era entender qual era o sentimento mais frequente e as palavras mais frequentes por sentimentos presentes nos tweets coletados, para assim termos uma ideia se o *trailer* teve uma boa recepção. Estes tweets coletados chamaremos de estudo de caso 1.

Para nosso segundo assunto foram coletados tweets relacionados ao Covid-19, com o objetivo de termos uma ideia do que os usuários brasileiros sentem em relação ao Covid-19 após mais de 1 ano de pandemia e com campanha de vacinação em andamento. Para está coleta utilizamos a hashtag *covid19*. Estes tweets coletados chamaremos de estudo de caso 2. Em ambos os assuntos coletamos 1200 tweets no dia 16 de Novembro de 2021.

Ao analisarmos como o SentimentalPeaks classificou os tweets, verificamos que 613 tweets foram classificados como positivos, 370 tweets como neutros e 217 como negativos. Mostrando que cerca de 51.08% de todos os tweets foram classificados como positivos, o que nesta amostra coletada mostra que o *Trailer* teve uma boa recepção do público. A tabela **XII** exibe a nuvem de palavra de todos os tweets classificados, na qual conseguimos notar que boa parte das palavras nos tweets em positivo estão relacionadas há nomes de personagens aguardados no filme como ‘tobey’ e ‘andrew’, atores que interpretaram o personagem do Homem-Aranha em outros filmes.

Ao analisarmos como o SentimentalPeaks classificou os tweets, temos 189 tweets classificados como positivos, 250 tweets classificados como neutros e 761 classificados como negativos, mostrando que cerca de 63.41% de todos os tweets foram classificados como negativos, mostrando que o tema Covid-19 é visto como algo negativo. A tabela **XIII** exibe a nuvem de palavra de todos os tweets classificados, na qual

⁸<https://www.adorocinema.com/noticias/filmes/noticia-161057/>



Tabela XII
NUVEM DE PALAVRAS DO ESTUDO DE CASO 1

conseguimos notar que boa parte das palavras nos tweets em negativos são: dose, reforço, Vacina, que acreditamos está relacionado ao anúncio de uma dose de reforço da vacina do Covid-19. Já a figura 9 exibe o uso da visualização do SentimentalPeaks e como podemos observar a um número muito maior de tweets negativos do que positivos ou neutros e tweets com mais retweets são também tweets negativos.

C. Uso com o PeakVis

Para demonstrar o uso do SentimentalPeaks com o PeakVis realizamos um experimento utilizando tweets coletados com a *hashtag* 'amordemae', referentes ao último capítulo de uma novela brasileira chamada 'amor de mãe', e a gravação em video do capítulo transmitido. Nossso objetivo com este experimento é identificar qual o sentimento mais presente em momentos de maior pico de tweets durante a transmissão do capítulo e o que estava sendo mais retuitado nestes momentos.

Ao observarmos as figuras 10 e 11 notamos uma maior quantidade de tweets em momentos que estão ocorrendo os intervalos na novela, ou seja, os momentos nos quais podemos definir intervalos como momentos no quais são exibidos anúncios de outras empresas. Na figura 10 realizamos o filtro dos tweets exibidos por uma das palavras mais frequentes: 'thelma'. Em um dos momentos de pico observamos uma maior quantidade de tweets negativos com maior quantidade de retweets. Tais tweets também apresentaram uma grande ocorrência de palavras como 'morrer' e 'aunerisma'.

Na figura 11 capturamos outro momento de pico, mas não realizamos nenhum filtro. Podeemos observar neste caso uma maior quantidade de tweets negativos, contendo frases como 'thelma colocando a culpa na aunerisma' e 'sinto falta do casseta planeta fazendo paródia de novelas'. Já nos tweets positivos observamos frases como 'a Adriana esteves segue sendo a rainha do pop' e 'que atuação impecável de adriana esteves'. Ou seja, é feita uma menção positiva à atriz.

IX. CONCLUSÃO E TRABALHOS FUTUROS

O nosso trabalho apresentou o desenvolvimento de um modelo de aprendizado de máquina, combinando 3 algoritmos que foram treinados e testados de forma individual, para classificar sentimentos em tweets e uma visualização para tweets classificados. A utilização de dados de redes sociais remete a dificuldades extras, pela não estruturação dos textos e muitas vezes pela falta de formalismo, além de uso de

vocabulário próprio [10]. Desta forma, para evitarmos problemas conhecidos como *Garbage in, Garbage out*, no qual uma máquina produz resultados ruins a partir de dados de entrada ruins, é necessário utilizar técnicas corretas de pré-processamento textual Medina [10].

A primeira hipótese não conseguimos validar, mesmo utilizando dois conjuntos de dados diferentes com aumento na quantidade de tweets. Ao analisarmos o experimento na Seção VIII e compararmos as tabela VIII-A e VIII-B com os resultados do trabalho de Medina [10], notamos que o conjunto de dados com as hashtags: economia, politica, saúde, entretenimento e educação do segundo semestre de 2021 obtiveram resultados inferiores as mesmas hashtags do primeiro semestre de 2021.

Desta forma, observamos que alguns assuntos são sensíveis à mudanças de sentimentos de forma mais rápida que outros, gerando ruídos nos algoritmos no processo de treinamento. Por exemplo, no assunto economia observamos que nos dados do primeiro semestre de 2021 a palavra “bolsonaro” apareceu com bastante frequência em tweets neutros, mas nos dados do segundo semestre esta mesma palavra encontra-se com bastante frequência em tweets de sentimento negativo. A tabela XIV exibe a comparação dos conjuntos de dados do primeiro e segundo semestre através de nuvens de palavras na categoria economia. Podemos notar também que a quantidade de *Hashtags* influencia em como os algoritmos classificam novos dados, pois, ao analisarmos o experimento na Seção VIII as tabelas VIII-A e VIII-B mostram que somente o aumento dos tweets por *hashtags*, obteve resultados inferiores ao trabalho de Medina [10]. Também o conjunto de dados 1 que contém um maior número de *Hashtags* obteve resultados piores do que o conjunto de dados 2.

Já a segunda e terceira hipótese conseguimos validar conforme observado na tabela **VIII-A**, a qual mostra que nosso algoritmo combinado obteve acurácia acima de 70% e melhores resultados do que os algoritmos sozinhos. Referente a nossa visualização de tweets conseguimos observar que a mesma é capaz de auxiliar na extração de *insights* úteis para análise de tweets.

Ainda há bastante espaço para melhoria dos resultados. Observamos que ao realizar o balanceamento dos dados com a técnica *Undersampling* perdemos cerca de 37% dos tweets o que significa que perdemos informações que poderiam ajudar no treinamento dos algoritmos. Para trabalhos futuros, poderemos resolver este problema coletando e anotando novos

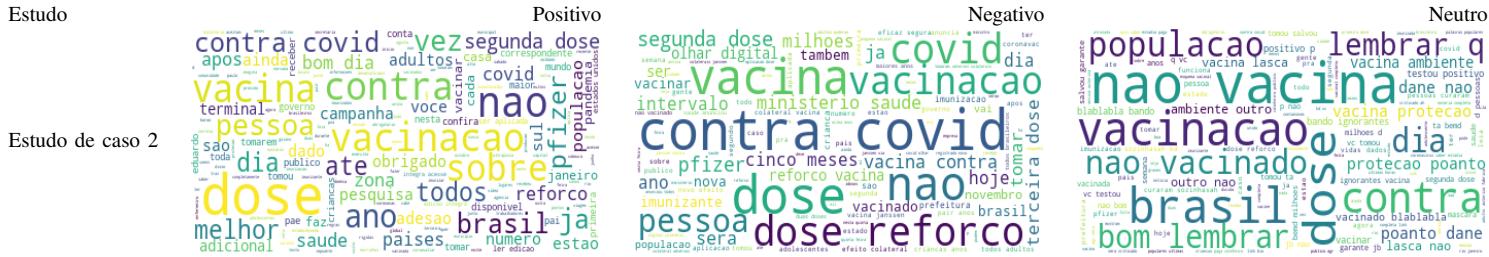


Tabela XIII
NUVEM DE PALAVRAS DO ESTUDO DE CASO 2

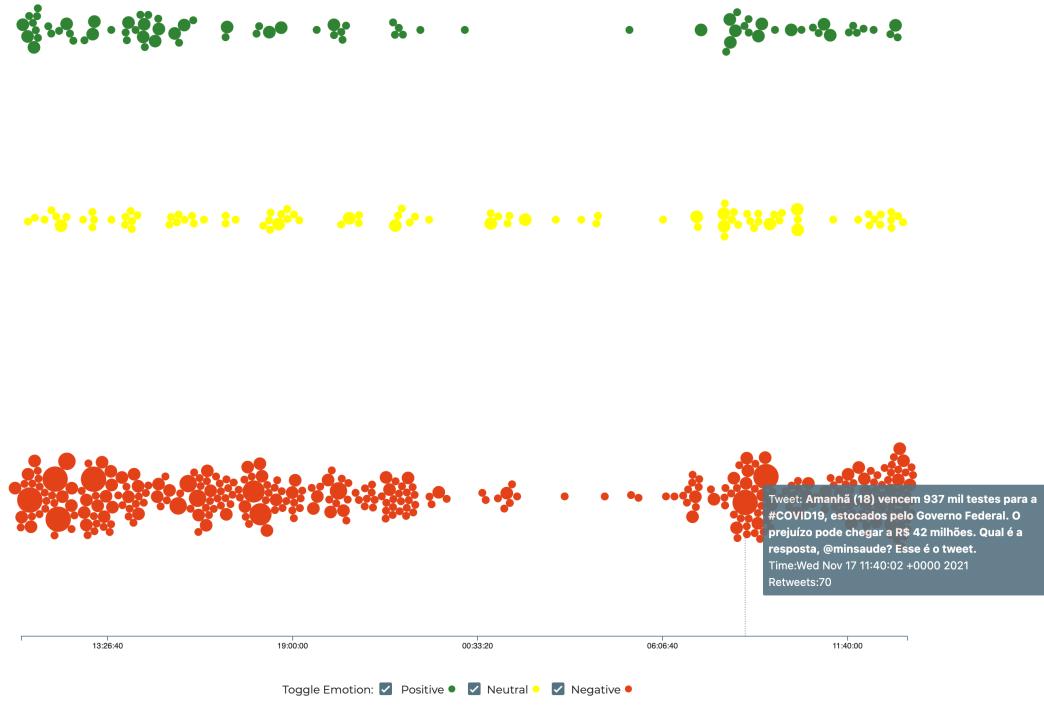


Fig. 9. Visualização de tweets estudo de caso 2

dados e incluindo estes dados no conjunto de dados de forma balanceada, ou seja, se tivermos maior quantidade de um determinado sentimento só incluiremos novos dados deste sentimento caso os demais estejam平衡ados. Outra forma de melhoria, seria aplicar este mesmo processo de forma mais profunda ao aplicar esta técnica a nível de *Hashtag*, ou seja, ter os dados平衡ados por *Hashtag* e sentimentos.

Outra abordagem que poderíamos testar com o nosso algoritmo seria usar modelos que realizam as classificações em nível de hierarquia. Assim, caso um modelo classifique como positivo um tweet, outro modelo treinado para classificar em positivo realiza esta classificação, caso não tenha classificado como positivo outro modelo realiza a classificação com o objetivo de utilizar a capacidade de todos os modelos de aprendizado de máquina desenvolvidos.

Além disso, novas abordagens de classificação mais modernas podem ser utilizadas, como Redes Neurais e Aprendizado Profundo(*Deep Learning*). Estas técnicas já utilizadas em outros trabalhos e apresentaram resultados de 74% [19] e

94% [3] de acurácia.

Como contribuição, todo o código desenvolvido e todos os conjuntos de dados composto por 1.700 tweets em conjunto com os 2.000 tweets coletados no trabalho [10] anotados manualmente, estão disponibilizados na plataforma Github para utilização livre. O modelo de aprendizado de máquina e o conjunto de dados estão em um repositório⁹, e a visualização e integração do modelo com o PeakVis estão em outro¹⁰.

REFERÊNCIAS

- [1] Erikson Aguiar, Bruno Faiçal, Jó Ueyama, Glauco Carlos Silva, and André Menolli. 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*, pages 393–406, Porto Alegre, RS, Brasil, 2018. SBC.
- [2] Larissa Britto. A text analysis approach for cooking recipe classification based on brazilian portuguese documents. In *Anais do XVII Encontro Nacional de Inteligência Artificial e Computacional*, 2019.

⁹<https://github.com/gustavohd18/TCC-analise-sentimentos>

¹⁰<https://github.com/DAVINTLAB/Peakvis>

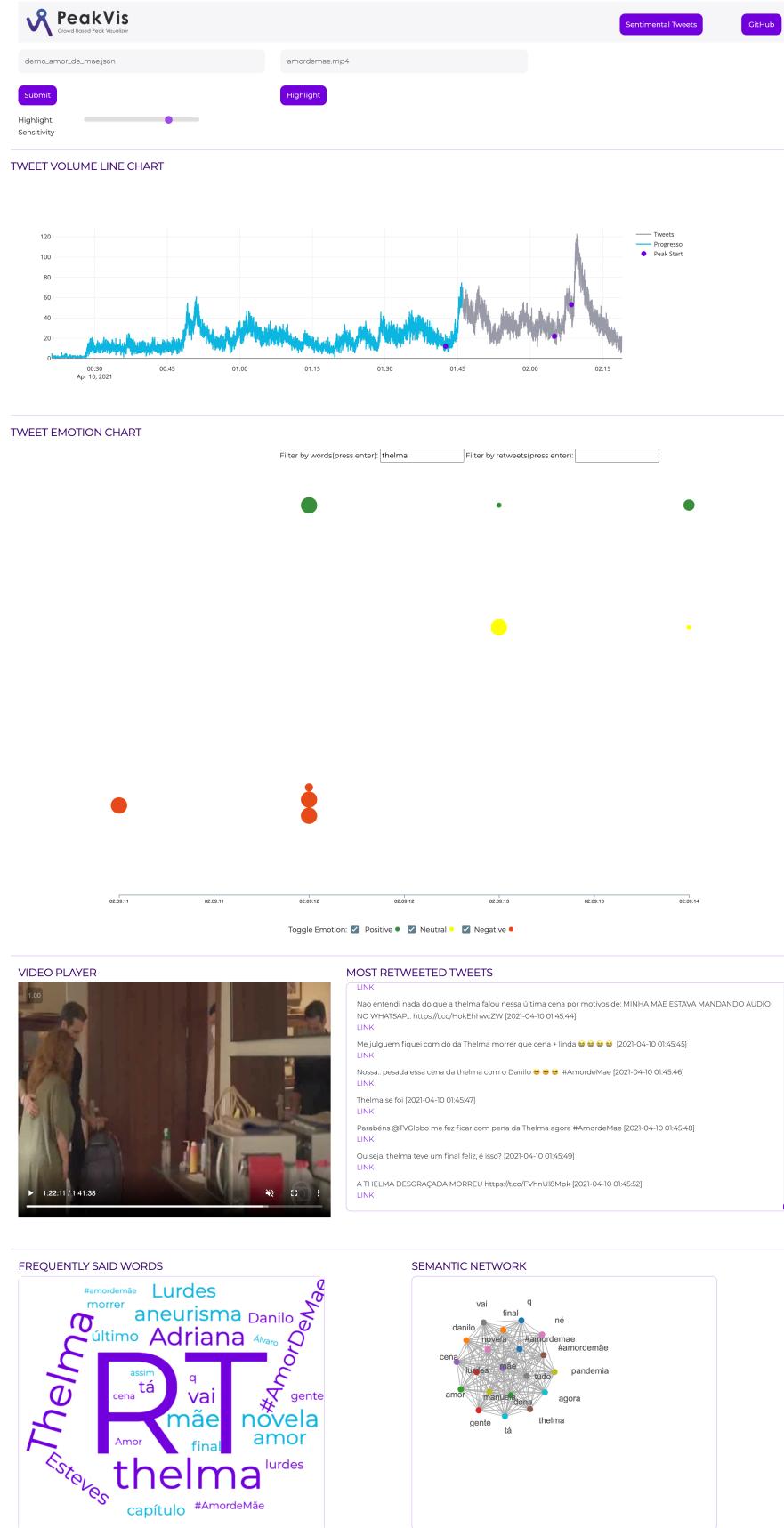


Fig. 10. SentimentPeaks com Peaks tweets filtrados.

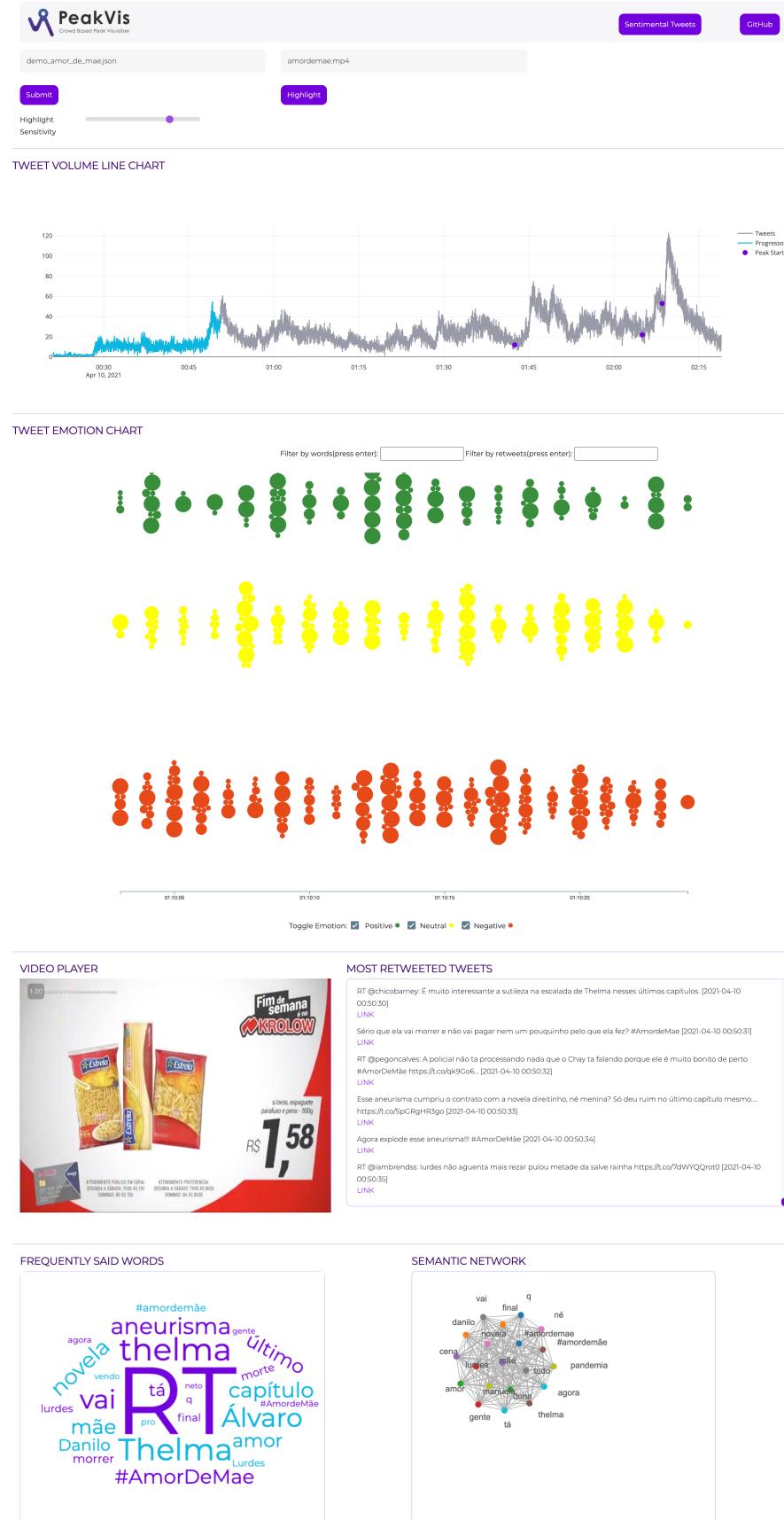


Fig. 11. SentimentalPeaks com Peaks sem filtros de tweets.



Tabela XIV
COMPARATIVO TWEETS PRIMEIRO E SEGUNDO SEMESTRE NUVEM DE PALAVRAS ECONOMIA

- [3] Golcalves Alexandre e Todesco José Crescencio Marcio. Um processo de classificação de texto: Análise de sentimento das opiniões no tripadvisor® sobre a atração oktoberfest blumenau. 2020.
- [4] Shihab Elbagir and Jing Yang. Sentiment analysis of twitter data using machine learning techniques and scikit-learn. In *Proceedings of the 2018 International Conference on Algorithms, Computing and Artificial Intelligence*, ACAI 2018, New York, NY, USA, 2018. Association for Computing Machinery.
- [5] Shihab Elbagir and Jing Yang. Sentiment analysis of twitter data using machine learning techniques and scikit-learn. In *Proceedings of the 2018 International Conference on Algorithms, Computing and Artificial Intelligence*, ACAI 2018, New York, NY, USA, 2018. Association for Computing Machinery.
- [6] Katti Faceli. *Inteligência artificial uma abordagem de Aprendizado de Máquina*. LTC, 1 edition, 8 2011.
- [7] Nann Hwan Khun and Hninn Aye Thant. Visualization of twitter sentiment during the period of us banned huawei. In *2019 International Conference on Advanced Information Technologies (ICAIT)*, pages 274–279, 2019.
- [8] Nann Hwan Khun and Hninn Aye Thant. Visualization of twitter sentiment during the period of us banned huawei. In *2019 International Conference on Advanced Information Technologies (ICAIT)*, pages 274–279, 2019.
- [9] Wei-Chao Lin, Chih-Fong Tsai, Ya-Han Hu, and Jing-Shang Jhang. Clustering-based undersampling in class-imbalanced data. *Information Sciences*, 409-410:17–26, 2017.
- [10] Thiago Pizzio Medina. Um estudo sobre classificação de tweets em português utilizando aprendizado de máquina, 2021. Trabalho de Conclusão de Curso, PUCRS-Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre, Brazil.
- [11] Vijayarani Mohan. Preprocessing techniques for text mining - an overview. 02 2015.
- [12] Saidah Saad and Bilal Saberi. Sentiment analysis or opinion mining: A review. *International Journal on Advanced Science, Engineering and Information Technology*, 7:1660, 10 2017.
- [13] Jayakumar Sadhasivam, Ramesh Babu, and Senthil Jayavel. Survey of various algorithms used in twitter for sentiment analysis. *International Journal of Engineering Trends and Technology*, 68:59–65, 04 2020.
- [14] Pedro Henrique M. Sanvido, Gabriela B. Kurtz, Carlos R. G. Teixeira, Pedro P. Wagner, Lorenzo P. Leuck, Milene S. Silveira, Roberto Tietz-mann, and Isabel H. Manssour. Peakvis: a visual analysis tool for social network data and video broadcasts. *COMPSSAC*, 2021.
- [15] Renata Vieira and Lucelene Lopes. Processamento de linguagem natural e o tratamento computacional de linguagens científicas. *EM CORPORA*, page 183, 2010.
- [16] Ribeiro A. Batista N. A. Vilela, R. B. Nuvem de palavras como ferramenta de análise de conteúdo: Uma aplicação aos desafios do mestrado profissional em ensino na saúde. pages 29–36, 2020.
- [17] Yun Wan and Qigang Gao. An ensemble sentiment classification system of twitter data for airline services analysis. In *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*, pages 1318–1325, 2015.
- [18] Florence Ying Wang, Arnaud Sallaberry, Karsten Klein, Masahiro Takatsuka, and Mathieu Roche. Senticompass: Interactive visualization for exploring and comparing the sentiments of time-varying twitter data. In *2015 IEEE Pacific Visualization Symposium (PacificVis)*, pages 129–133, 2015.
- [19] Jônatas Wehrmann, Willian Becker, and Rodrigo Barros. A multi-task neural network for multilingual sentiment classification and language detection on twitter. 01 2018.
- [20] Nikhil Yadav, Omkar Kudale, Srishti Gupta, Aditi Rao, and Ajitkumar Shitole. Twitter sentiment analysis using supervised machine learning. 04 2020.
- [21] Jian Zhao, Liang Gou, Fei Wang, and Michelle Zhou. Pearl: An interactive visual analytic tool for understanding personal emotion style derived from social media. In *2014 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 203–212, 2014.
- [22] David Zimbra, Ahmed Abbas, Daniel Zeng, and Hsinchun Chen. The state-of-the-art in twitter sentiment analysis: A review and benchmark evaluation. *ACM Trans. Manage. Inf. Syst.*, 9(2), August 2018.