Mineração de Texto para a Análise do Perfil Emocional de Usuários de Jogo Empático

Leonardo Dias Martins
Universidade Federal do Pará
leodm89@gmail.com

Fabíola Pantoja Oliveira Araújo
Universidade Federal do Pará
fpoliveira@ufpa.br

ABSTRACT

Daily, a large amount of data circulates on the Internet, producing a lot of information in the form of images, videos and texts. Then, it is necessary to analyze and extract these information automatically. Therefore, this work presents a case study that applies text mining to extract the emotional and sentimental profiles from the comments of the Last Day of June game users, where the results and the information extracted from the analysis of sentiments were presented. Three classification algorithms were used: Naive Bayes, Support Vector Machine (SVM) and K-Nearest Neighbors (KNN) to predict the class of elements according to the emotions or feelings identified in the comments analysis. As a result, SVM with radial kernel was the one with the best accuracy, with 79%, followed by KNN with 3 closest neighbors, with 75%, and finally, Naive Bayes, with 62%.

KEYWORDS

Computação afetiva, Mineração de texto, Jogos empáticos.

1 INTRODUÇÃO

Atualmente, existem várias alternativas tecnológicas para ajudar e entreter a população. Entre uma dessas, estão os jogos digitais desenvolvidos para atenderem um vasto e exigente mercado consumidor. Com isso, os jogos digitais estão tornando-se cada vez mais complexos e realistas, proporcionando uma imersão maior dos usuários com o jogo, por meio de recursos gráficos, efeitos sonoros avançados e aparatos físicos como os óculos de realidade virtual, sensores para reconhecimento de movimentos, fones de ouvidos preparados para melhorar a experiência sonora entre outros.

Nos jogos empáticos, que tem por objetivo simular aos jogadores experiências e sensações vividas por outras pessoas e que colaborem para que, por meio de uma experiência virtual, o jogador possa imaginar-se em situações de dificuldades e como lidar com elas, pode ser interessante realizar a análise emocional nas mensagens enviadas pelos jogadores, pois é possível identificar quais sentimentos foram mais estimulados durante a interação com o jogo, podendo verificar as emoções predominantes, examinar o impacto que a narrativa causou na forma de pensar do jogador entre outras.

Existem vários métodos que podem ser utilizados na coleta e tratamento de dados textuais durante a análise afetiva dos comentários de usuários de jogos digitais, uma delas é a mineração de texto, que pode ser feita de forma manual ou automatizada. Em alguns casos, em que a quantidade de dados coletados e analisados é pequena, torna-se possível o tratamento textual ser feito de forma manual, porém na maioria das situações a quantidade de dados a ser estudada é grande, de forma que fica inviável ser feita dessa maneira, demandando bastante tempo para isso. As técnicas automatizadas são as mais utilizadas e recomendadas, por utilizarem ferramentas computacionais que realizam o processamento e tratamento dos dados textuais de maneira ágil e automática, sendo necessário apenas que o usuário configure a ferramenta para realizar o trabalho, de acordo com o seu objetivo desejado.

Considerando todo o contexto exposto, este trabalho apresenta um estudo de caso que utiliza a mineração de texto para identificar padrões emocionais e sentimentais a partir da análise automática de comentários dos usuários do jogo "*Last Day of June*". Este artigo está organizado da seguinte forma: na Seção 2 é apresentado o referencial teórico assim como os trabalhos correlatos, a metodologia do trabalho é detalhada na Seção 3, a análise dos resultados é realizada na Seção 4 e as considerações finais na Seção 5.

2 REFERÊNCIAL TEÓRICO

Os jogos digitais têm um catálogo variado com diversos enredos e temáticas diferentes, em que o jogador interage com a história e em algumas vezes se imagina no lugar do personagem do jogo, tendo que tomar decisões para avançar e solucionar desafios que surgem durante o decorrer do jogo.

Durante a interação do usuário com o jogo, algumas sensações são envolvidas como sentimentos, emoções, imaginações entre outras, que são estimuladas por meio das imagens e sons exibidos. Os efeitos e a influência dos jogos digitais na forma de pensar e no comportamento dos jogadores, principalmente em jovens e adolescentes, é objeto de estudos e debates [1], portanto, fica claro a importância de analisar alguns elementos como comentários, comportamentos e atitudes dos usuários com o intuito de obter informações acerca dos efeitos causados pelos jogos em seus usuários.

A empatia está relacionada diretamente com as emoções e sentimentos, pois permite que pessoas possam entender o que

outras estão sentindo, se imaginando no lugar delas e, assim, compreender os sentimentos e emoções alheias [2].

Nos jogos empáticos o objetivo é fazer o usuário se imaginar no lugar do personagem da história, que passa por algum tipo de problema que existe na vida real, e mostrar como é a percepção desse problema pelo ponto de vista de quem está enfrentando-os [3]. Nos jogos empáticos também são apresentados desafios para o jogador concluir, com a finalidade de fazer o personagem encontrar formas de enfrentar e superar suas dificuldades [3].

Existem algumas formas de analisar as informações disponíveis na Internet referentes a usuários de determinado produto ou serviço, podendo ser de forma manual, em que os dados são coletados e analisados manualmente de forma individual ou em pequenos grupos ou de forma automática, quando são utilizadas técnicas computacionais automatizadas para fazer a análise de volume de dados maiores.

Uma das formas de análise de dados automatizadas é a mineração de dados, onde os dados são coletados e analisados com a ajuda de programas computacionais, que realizam ações no dado coletado com o objetivo de encontrar informações úteis [4]. Algumas das principais etapas da mineração de dados é o pré-processamento dos dados, em que são realizadas ações para adequar e padronizar os dados e, assim, facilitar a análise [4]. Outra tarefa importante é a eliminação de elementos que não tenham relevância para o objetivo da mineração [4].

Quando a finalidade é encontrar informações em dados textuais, é utilizada a mineração de texto a qual se assemelha à mineração de dados, em que o dado coletado (que nesse caso será somente texto) será preparado na etapa de pré-processamento do texto, onde são feitas as remoções de termos sem importância, como as *stopwords*, que são elementos sem valor semântico para o texto [4]. Outras ações podem ser realizadas para facilitar e otimizar os resultados, como a criação de tabelas de elementos organizados por ordem de frequência ou categorias, sendo que essas ações variam de acordo com a finalidade desejada [5].

A computação afetiva é a área em que se estuda maneiras de fazer o computador reconhecer padrões das emoções e sentimentos humanos, assim como também, de simular essas características de forma artificial para que a interação do ser humano com o computador seja mais eficaz [6].

Dessa maneira, Rosalind [7] afirma que o reconhecimento e simulação de características afetivas humanas em computadores são importantes para que eles se tornem máquinas mais inteligentes e proporcionem uma interação humano-computador mais natural.

Dentre um dos ramos da mineração de dados, existe a análise de sentimentos ou mineração de opiniões, a qual é a busca por padrões emocionais e sentimentais em textos, com o objetivo de compreender e reconhecer, a partir de comparações dos elementos dos textos com dados de emoções e sentimentos já classificados, as características afetivas do público analisado, com isso, colaborando com o melhor entendimento do que as pessoas escrevem e na tomada de decisões[8].

Muitas empresas já usam a análise de sentimentos para compreender melhor as emoções e sentimentos de seus clientes em relação a seus produtos e serviços, podendo assim, proporcionar alternativas mais específicas e próximas do que eles desejam [9].

2.1 Trabalhos correlatos

Alguns trabalhos já abordaram a utilização da mineração de texto para obter informações sobre determinados temas, como em [10], que aplica a mineração de texto para realizar a análise de sentimentos em letras de músicas. Santos et al. [11] analisaram mensagens de usuários de fóruns de discussão para encontrar informações relacionadas a alguns aspectos, como emoções e sentimentos, aspectos técnicos sobre o jogo, ao luto e a mortalidade. Já Aguiar et. al. [12] utilizaram algoritmos de classificação para realizar a análise de sentimento em mensagens de usuários inseridas em redes sociais e compararam os resultados dos seus desempenhos na tarefa de classificação dos dados. Rocha [13], realizou um estudo que utiliza algoritmos de classificação para classificar processos judiciais trabalhistas de acordo com os assuntos abordados nos processos. Foi feita uma análise em um grupo de 241 mil documentos do tipo Recursos Ordinários, com o objetivo de encontrar o assunto principal de cada um deles e assim classificá-los com a ajuda dos algoritmos de classificação.

A principal vantagem deste trabalho em relação aos trabalhos de Souza et al. [10] e Santos et al. [11] é a utilização de algoritmos de classificação para prever a classe dos elementos. O diferencial em relação ao trabalho de Aguiar et al. [12] é a utilização de múltiplas fontes de dados, já que nele foram utilizados dados de apenas uma fonte (*Twitter*). Em relação ao trabalho de Rocha [13], que também utiliza a mineração de texto e algoritmos de classificação para prever e classificar os elementos, a vantagem é a utilização do algoritmo de classificação K-Vizinhos mais próximos (*KNN*), que permite utilizar diferentes parâmetros de classificação, podendo adaptar-se melhor a variados tipos de dados.

3 METODOLOGIA

Neste artigo foi utilizado um estudo de caso em que foram coletados, de forma manual, comentários no idioma português do Brasil de usuários do jogo *Last Day of June*[1], dentro do *site* da plataforma de jogos *Steam*. Posteriormente, esses comentários foram tratados computacionalmente e analisados, utilizando técnicas de mineração de texto para encontrar padrões afetivos (emoções e sentimentos).

3.1 Procedimentos preparatórios

Nesta seção são descritos os procedimentos que foram realizados na análise dos sentimentos dos comentários dos usuários do jogo *Last Day of June,* desde a coleta dos comentários, processamento e finalização da análise.

3.1.1 Escolha do jogo: Foi realizada uma pesquisa em sites na Internet sobre alguns jogos empáticos, seus enredos foram

[1] https://store.steampowered.com/app/635320/Last_Day_of_June/

analisados para se chegar àquele que mais tivesse a capacidade de estimular variadas emoções em seus jogadores. Também foi verificado se o jogo escolhido tinha uma página com uma área dedicada para usuários expressarem suas opiniões por meio de comentários no idioma português do Brasil.

Seguindo os critérios descritos acima, o jogo escolhido foi o *Last Day of June* que narra a história de um casal, Carl e sua esposa June, que durante uma viagem de carro sofrem um acidente. Carl perde os movimentos das pernas, passando a usar cadeira de rodas, e June não consegue sobreviver ao acidente [14]. Durante o jogo, o papel do jogador é voltar até o dia do acidente e realizar ações para que o desfecho seja diferente, tentando evitar os efeitos apresentados no início da história.

3.1.2 Coleta dos comentários: Foram coletados 110 comentários de usuários do jogo na Internet, em fontes diferentes. A maior parte (93 comentários) foram da página localizada na plataforma de jogos Steam, na seção de análises, em que usuários podem deixar comentários sobre o que acharam do jogo. Os restantes (17 comentários) foram obtidos nos sites *Twitter*[2] e *Youtube*[3].

3.1.3 Preparação do ambiente: Para realizar a análise dos comentários coletados foi utilizado um computador com sistema operacional *Windows 10,* com os programas *RStudio*[4] (versão 1.1.456) e *R*[5] (versão 3.6) instalados, ambos para trabalhar na linguagem de programação R.

A linguagem de programação R foi escolhida por ser uma das mais populares (junto com *Python*) para se trabalhar com análise de dados [15] e também por possuir pacotes específicos para a análise de dados, como o "*tm"* - com funções voltadas para a mineração de texto e o "*syuzhet"* - para a análise de sentimentos.

3.1.4 Pré-processamento do texto: Nesta etapa são realizados alguns procedimentos com o objetivo de preparar o texto para a etapa da análise dos elementos dos comentários.

3.1.4.1 Leitura dos comentários: Os 110 comentários coletados dos usuários do jogo *Last Day of June* foram organizados e numerados em um arquivo (*dataset*[6]) no formato de texto simples (TXT). No programa *RStudio* foi criado um script, onde foi escrito todo o código responsável pelas etapas da análise afetiva dos comentários. Nesse script foi utilizada a função do *R* chamada "*read_lines"* para ler os comentários do arquivo de texto e passar para um arquivo dentro do *RStudio*, adequando-o para o restante dos procedimentos.

3.1.4.2 Tratamento do texto: Esta etapa é uma das mais importantes do pré-processamento do texto, pois é nela que termos irrelevantes para o trabalho são descartados, como *stopwords*, que é uma lista de termos sem relevância semântica para o texto como conectivos, pontos, números, símbolos especiais entre outros. Essas etapas podem variar de acordo com o objetivo do estudo [16].

Nos comentários utilizados neste artigo foram removidas as *stopwords,* presentes em uma lista já pronta encontrada na Internet, no site *GitHub*[7]. Outras palavras sem importância para o estudo foram removidas também, como os nomes dos personagens, pontuações e números, assim como espaços desnecessários presentes nos textos.

3.1.4.3 Transformação dos dados: Após a remoção dos termos irrelevantes, os dados foram transformados para tipos que facilitassem a análise dos elementos. Para isso, foi criado um documento do tipo matriz, com os elementos do texto organizados por ordem de frequência em que apareceram, e a função responsável por essa ação é a "*TermDocumentMatrix"*. Outra alteração feita foi a alocação dos termos e suas frequências em um outro documento no formato *dataframe,* o qual é um documento em forma de matriz também e que aceita caracteres de tipos diferentes e facilita na manipulação dos dados.

3.2 Identificação das emoções

Depois dos dados já terem sidos preparados, foi realizada a análise afetiva neles, em que os perfis emocionais e sentimentais são reconhecidos a partir da comparação dos elementos dos textos extraídos dos 110 comentários dos usuários com um dicionário de emoções e sentimentos presente na função "*get_nrc_sentiment"* do *R,* em que ela compara cada elemento com um dicionário (*NRC Emotion Lexicon*) que contém aproximadamente 14000 palavras de 105 idiomas, inclusive o português. Caso a palavra comparada esteja presente no dicionário, é verificada a qual emoção ela está associada, assim identificando sua relação com alguma emoção. As emoções presentes no dicionário *NRC* são: alegria, tristeza, raiva, medo, confiança, desgosto, surpresa e expectativa (*joy, sadness, anger, fear, trust, disgust, surprise e anticipation*). Além das 8 emoções citadas anteriormente, a função também identifica 2 sentimentos, positivo e negativo.

O processo de identificação das emoções e sentimentos inicia-se quando a função "*get_nrc_sentiment"* é utilizada para ler o documento com os elementos dos comentários, sendo feita a comparação de cada elemento com os presentes no dicionário *NRC,* assim, encontrando as relações dos elementos com as emoções e com os sentimentos. Após a comparação, a função salva um novo documento do tipo *dataframe* com as emoções e sentimentos divididos em colunas, uma para cada emoção e sentimento. Cada linha deste documento representa uma linha lida do texto, ou seja, um comentário. Como resultado se tem uma matriz com campos e valores referentes à quantidade de elementos em cada comentário, relativo a cada emoção (Tabela 1).

[2] https://twitter.com/
[3] https://www.youtube.com/
[4] https://www.rstudio.com
[5] https://www.r-project.org
[6] https://gist.github.com/L3oDM/0698da81342b59d0472db10e662eda17
[7] https://github.com/stopwords-iso/stopwords-pt

COMENTÁRIOS	RAIVA	EXPECTATIVA	DESGOSTO	MEDO	ALEGRIA	TRISTEZA	SURPRESA	CONFIANÇA	EMOÇÃO PREDOMINANTE	POLARIDADE
2:Acabei de concluir o game	25	0	0	25	0	50	0	0	TRISTEZA	NEUTRO
3:Jogo muito bom, com uma	0	25	0	0	50	0	0	25	ALEGRIA	POSITIVO
4: Isso não é um jogo, mas	100	0	0	0	0	0	0	0	RAIVA	NEGATIVO
10: Um dos jogos mais boni	12	12	8	16	8	24	8	12	TRISTEZA	NEGATIVO
13: Compre o jogo. Por quê?	0	0	0	0	0	100	0	0	TRISTEZA	NEGATIVO
18: Que jogo! Com certeza vale	25	0	0	25	0	50	0	0	TRISTEZA	NEGATIVO
20: Um dos jogos mais tocantes	12.50	12.50	0	12.50	12.50	37.50	0	12.50	TRISTEZA	NEGATIVO

Figura 1 - Trecho do *dataframe* DF.Geral

Tabela 1 – *Dataframe* Comentários x Emoções.

Comentários	Raiva	Expectativa	Desgosto	Medo	Alegria	Tristeza	Surpresa	Confiança	Negativo	Positivo
1	0	0	0	0	1	0	0	1	2	1
2	1	0	0	1	0	2	0	0	1	1
3	0	1	0	0	2	0	0	1	0	4
4	1	0	0	0	0	0	0	0	1	0
5	3	6	2	3	3	2	4	1	9	7
6	0	0	0	0	0	0	0	0	0	0
7	1	0	1	1	1	0	1	1	2	1

Foram realizadas algumas operações para melhorar a visualização e organização dos dados e também, reunir o máximo de informações em um único *dataframe*. Primeiramente, as colunas referentes às emoções e sentimentos foram renomeadas para os nomes em português do Brasil. Os comentários foram adicionados em uma nova coluna dentro do *dataframe* do resultado da função "*get_nrc_sentiment*". A partir dos valores em cada coluna, foi calculada a porcentagem de cada emoção em cada um dos comentários. Foi adicionada também uma coluna denominada "EMOÇÃO PREDOMINANTE" que exibe a emoção com maior presença no comentário correspondente. As linhas com comentários sem emoção predominante foram descartadas. Por fim, foi adicionada a coluna "POLARIDADE", na qual é apresentada se o comentário tem o sentimento predominante positivo, neutro ou negativo. Na Figura 1 é possível visualizar uma parte do *dataframe* (nomeado como DF.Geral) resultante após as modificações feitas.

3.3 Escolha dos algoritmos de classificação

Os algoritmos de classificação são usados para prever a classe dos dados baseando-se no aprendizado realizado pelos algoritmos durante a etapa de treino. As características de todos os dados classificados e não-classificados são agrupadas nas categorias que mais se assemelham e, consequentemente, possuem mais

características em comum com as classes existentes nos dados classificados [17].

Três algoritmos foram escolhidos para realizar a classificação dos dados neste trabalho: *Naive Bayes*, Máquina de Vetor de Suporte *(SVM – Support Vector Machine*) e K-Vizinhos mais próximos *(KNN – K-Nearest Neighbor*). A escolha desses três algoritmos foi baseada em pesquisas em outros trabalhos que utilizaram algoritmos de classificação na área de ciência de dados, como em [12], [18] e [19].

4 RESULTADOS

Depois de realizadas as etapas de pré-processamento e execução dos algoritmos de classificação (mineração de texto), foi feita a análise dos comentários e classificação dos termos de acordo com as 8 emoções e 2 sentimentos. Esta seção apresenta os resultados obtidos da análise afetiva.

4.1 Porcentagem das emoções

O gráfico da porcentagem das emoções foi gerado a partir da função do *R* chamada "*pie*". Os dados utilizados para calcular a porcentagem de cada emoção foram os valores de cada linha referentes às emoções identificadas nos elementos dos comentários. Na Figura 2 é apresentado o gráfico com a porcentagem de cada uma das 8 emoções no qual é possível perceber que a tristeza foi a emoção mais encontrada entre os 110 comentários, com 20.8% do total, seguida da alegria com 17.1%, sendo o desgosto a emoção menos detectada, com 3.5%.

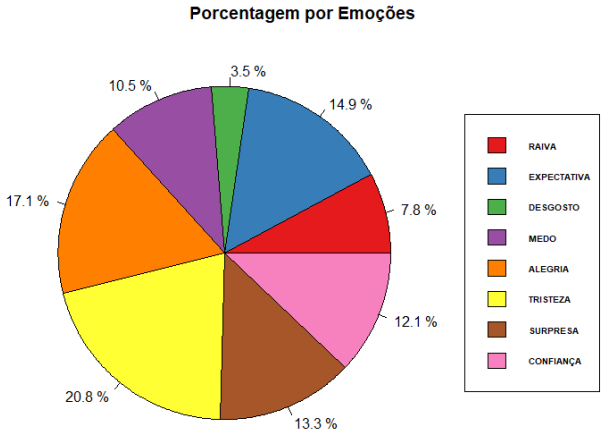

Figura 2 - Gráfico Pizza - Porcentagem das emoções.

Na Figura 2 é possível perceber que a tristeza foi a emoção mais encontrada entre os 110 comentários, com 20.8% do total, seguida da alegria com 17.1%, a emoção menos identificada foi o desgosto, com 3.5%.

4.2 Porcentagem dos sentimentos

O gráfico dos sentimentos foi gerado da mesma maneira que o das emoções, porém somente os valores correspondentes às colunas dos sentimentos foram utilizados para realizar o cálculo das porcentagens. Na Figura 3 é possível notar que os sentimentos positivo e negativo apresentaram percentuais iguais, em que cada um teve 50% de representação em relação ao total. A explicação para o equilíbrio entre os 2 sentimentos é a grande quantidade de comentários elogiando o jogo, os quais são classificados como positivos. Os acontecimentos no jogo, como o acidente de carro que provocou a morte de um dos personagens, e também o teor triste da história, colaboraram para a classificação do sentimento negativo.

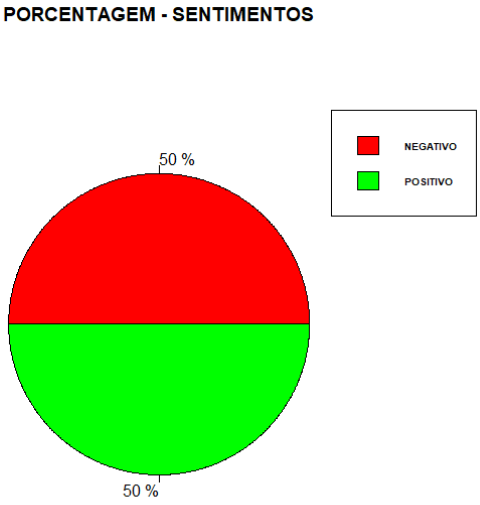

Figura 3 - Gráfico em pizza - Porcentagem dos Sentimentos.

4.3 Frequência das palavras por emoções

Nesta seção é apresentado o gráfico de barra com os valores de vezes que um elemento foi relacionado a uma ou mais emoções. Cada barra representa uma emoção, onde são exibidos no topo de cada barra o valor correspondente ao número de elementos identificados de acordo com a emoção relacionada.

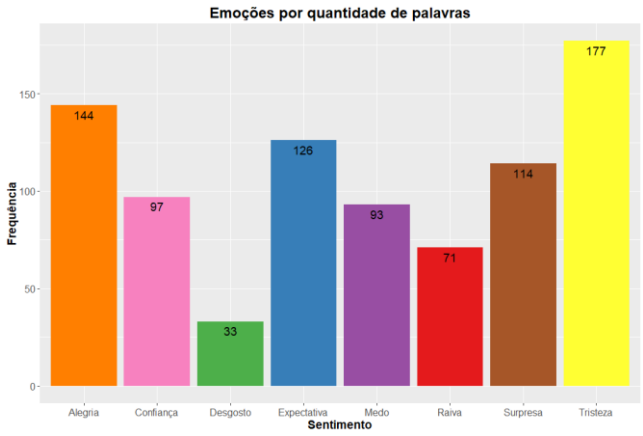

Figura 4 - Gráfico de barras da frequência dos sentimentos.

É possível notar na Figura 4 que a emoção mais predominante nos comentários foi a tristeza, com 177 elementos, seguida pela alegria, com 144 e da expectativa com 126. A maioria dos comentários são relacionados a sentimentos dos usuários em relação à história do jogo, alguns demonstrando alegria e satisfação, outros fazendo referência à tristeza sentida no momento em que ocorre a morte de uma personagem.

4.4 Diagrama de caixas ou boxplot das emoções

O diagrama de caixas *(boxplot)* é dividido em 5 regiões: a mediana, dois quartis que concentram 50% dos valores totais, um deles com valores maiores que a mediana e o outro com valores menores. Há também duas linhas, a superior e inferior, localizadas fora dos quartis, que representam os limites máximos de valores. Acima ou abaixo dessas duas linhas, são considerados valores discrepantes ou também chamados de *outliers*[20,21].

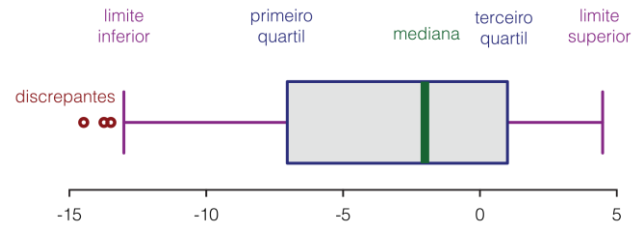

Figura 5 - Elementos de um *Boxplot*.

A Figura 6 mostra o *boxplot* das 8 emoções identificadas nos comentários. Nela, é possível identificar que a caixa maior é a que representa a emoção tristeza, pois é a emoção com mais elementos referentes a ela e a mais encontrada nos comentários, seguida respectivamente pela, alegria e expectativa. Em relação à mediana, somente a emoção alegria apresentou elementos com valores abaixo da mediana. A região dos limites inferiores de todas as emoções não apresentou elementos, devido os valores apresentarem a quantidade de elementos identificados por emoção, sendo o mínimo zero, representando que não foi encontrado elemento referente à emoção.

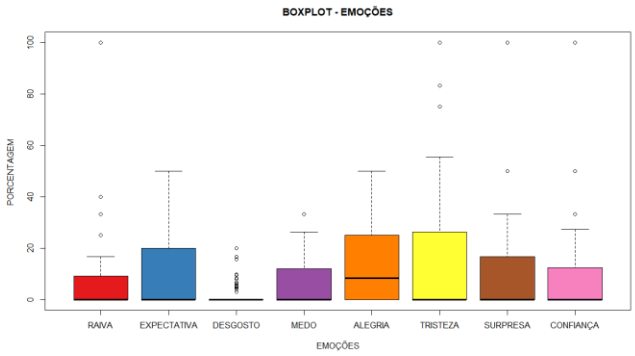

Figura 6 - *Boxplot* das 8 emoções.

Na região dos elementos discrepantes, não houve a presença de elementos abaixo dos limites inferiores, pelo mesmo motivo de a função não utilizar valores negativos, já em relação aos limites superiores, somente as emoções expectativa e alegria não apresentaram *outliers*, pois todos seus elementos apresentaram valores dentro dos limites. Desgosto foi a emoção com mais *outliers* superiores, pois como ela teve poucas ocorrências, isso colaborou para que sua mediana, seus quartis e bigodes, ficassem localizados na escala do valor zero. Consequência disso é a baixa quantidade de elementos com valores referentes à identificação do desgosto nos comentários, ficando seus elementos acima do limite superior, posicionados na região dos valores discrepantes.

As emoções raiva, tristeza, surpresa e confiança foram aquelas com *outliers* mais distantes dos limites superiores, pois elas tiveram 100% ou percentual próximo disso de identificação em um único comentário.

4.5 Correlação das emoções

O gráfico das correlações das emoções apresenta informações acerca das similaridades entre as emoções por meio de um mapa de calor com vários quadrantes, que possuem valores que variam em uma escala de 0.4 a 1.0, em que 0.4 representa o menor nível de correlação entre duas emoções e 1.0, o maior. Cada quadrante possui uma cor específica, em que os que possuem a cor menos intensa são os quadrantes referentes a emoções com menor grau de correlação entre si. Já aqueles com cores mais intensas são os que representam as emoções com maior grau de correlação mútua. A função do R utilizada para gerar o gráfico foi a "cor". Na Figura 7, é possível visualizar o gráfico com o mapa de calor das correlações entre as emoções.

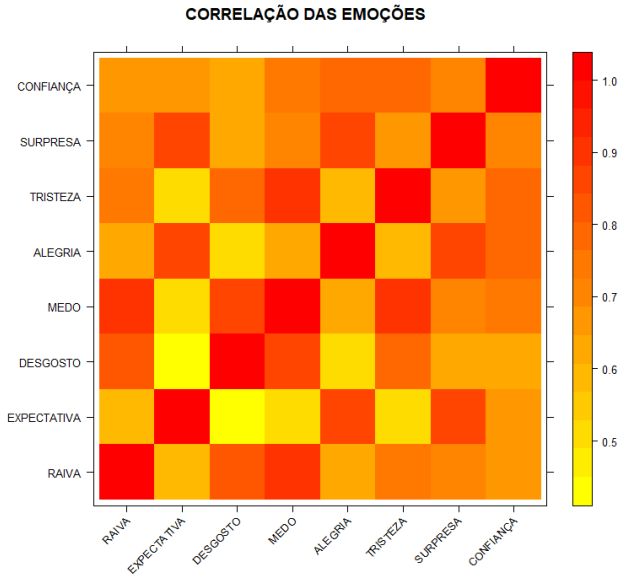

Figura 7 - Mapa de calor da correlação entre emoções.

Na referida figura é possível notar a similaridade entre algumas emoções, pois a cor é mais intensa nos quadrantes que representam esses pares de emoções, tais como entre surpresa-expectativa, surpresa-alegria, tristeza-medo, alegria-expectativa, medo- raiva, medo-desgosto, desgosto-raiva e desgosto-tristeza. Por outro lado, nos quadrantes com cores menos intensas ou mais claras, encontram-se as emoções com menor grau de correlação entre si, como é o caso de tristeza-expectativa, alegria-desgosto, medo-expectativa e expectativa-desgosto, ou seja, são os pares de emoções com menos similaridades, aparentando pertencerem a sentimentos opostos.

4.6 Resultados dos algoritmos de classificação

Foram utilizados três algoritmos para prever a classificação dos sentimentos (positivo e negativo) a partir dos dados gerados anteriormente pela função *get_nrc_sentiment*. São eles: o *Naive Bayes*, Máquina de Vetor de Suporte (*SVM*) e o K-Vizinhos mais próximos *(KNN*).

O *Naive Bayes* calcula a probabilidade de um evento ocorrer analisando a ocorrência de eventos anteriores, ou seja, é usada a probabilidade para calcular a chance de um dado não-classificado ser de determinada classe baseando-se nas classes de dados já classificados anteriormente [22]. O *SVM* utiliza uma separação dos dados em planos, chamados de hiperplanos, de acordo com suas classes. A classificação do elemento é feita analisando sua posição no hiperplano e os hiperplanos próximos, baseando-se nas classes dos elementos presentes nos hiperplanos mais próximos ou do seu hiperplano [23]. No algoritmo *KNN*, inicialmente todos os elementos são organizados em um espaço calculado pelo algoritmo. Após isso, analisa-se a posição do elemento a ser classificado e as posições dos seus vizinhos mais próximos. A classe predominante dos vizinhos mais próximos será aquela atribuída ao elemento a ser classificado [23].

Como resultado, o *Naive Bayes* obteve 45% de eficácia na tarefa de classificação dos sentimentos a partir dos dados dos sentimentos gerados pela função "*get_nrc_sentiment"*, porém com a utilização de estimativa de densidade de *kernel (KDE)*, esse percentual subiu para 62%. De acordo com Pierson [24]: "A estimativa de densidade do *kernel* (KDE) é um método de suavização; coloca um *kernel* – uma função de ponderação que serve para quantificar a densidade – em cada ponto de dados no conjunto e soma os *kernels* para gerar uma estimativa de densidade do *kernel* para a região em geral."

No algoritmo *SVM* tem como utilizar a opção de *kernel* para definir como os hiperplanos e os dados serão organizados no momento da classificação dos elementos. Os *kernels* utilizados nesse artigo foram: *linear* (Padrão), *radial* (*RBF – Radial Basis Functions*), *polynomial* e *sigmoid*. A utilização do *kernel radial* foi o que obteve melhor precisão, com 79%, seguido pelo *linear* com 68%, depois pelo *sigmoid*, com 54%, e por último (menos preciso), o *polynomial* com 34% de precisão.

No algoritmo *KNN* pode ser definida a quantidade de vizinhos analisados do elemento que será classificado. Neste trabalho foram utilizados valores de 3 e 5 vizinhos mais próximos, pois foram os valores com melhores resultados após alguns testes feitos. Na opção de analisar os 3 vizinhos mais próximos, o algoritmo obteve a precisão de 75%. Já com a análise dos 5 vizinhos mais próximos, o percentual de precisão caiu para 72%.

Após os resultados das análises feitas da utilização de algoritmos de classificação para prever a classe relativa aos sentimentos dos elementos, é possível perceber que o algoritmo *SVM* com a utilização do *kernel radial* foi o que obteve melhor resultado, com 79%. O menos preciso também foi o *SVM*, mas com a utilização do *kernel polynomial*, com apenas 34%.

Na Tabela 2 são apresentados todos os dados das porcentagens de precisão de cada algoritmo.

Tabela 2 - Precisão dos algoritmos de classificação.

Algoritmo	Precisão
SVM (*kernel Radial).*	79%
KNN (com 3 vizinhos mais próximos).	75%
KNN (com 5 vizinhos mais próximos).	72%
SVM (*kernel Linear).*	68%
Naive Bayes (com *KDE*).	62%
SVM (*kernel Sigmoid*).	54%
Naive Bayes.	45%
SVM (*kernel Polynomial*).	34%

Neste trabalho, o objetivo do uso do aprendizado de máquina através dos algoritmos de classificação foi prever os padrões emotivos de dados ainda não classificados, a partir do resultado do treinamento feito com os dados classificados nas etapas anteriores. Assim, foi possível prever e classificar, somente com o uso dos algoritmos de classificação, um grupo de dados que ainda não tinha sido classificado.

5 CONSIDERAÇÕES FINAIS

Este trabalho apresentou um estudo de caso que utiliza a mineração de texto para realizar a análise afetiva em comentários de usuários do jogo digital empático *Last Day of June*. Foi possível identificar padrões afetivos (emoções e sentimentos) a partir da análise dos comentários de usuários. Com isso, foi possível perceber a importância de conhecimentos que podem ser obtidos a partir da análise dos textos, podendo ser aplicado a outros contextos como as redes sociais, por exemplo, para detectar comportamentos ou emoções dos usuários.

A utilização da mineração de texto possibilitou a identificação de 8 emoções e 2 sentimentos presentes nos comentários analisados e uso de algoritmos de classificação para prever os sentimentos presentes nos comentários.

A partir dos resultados obtidos, conclui-se que a emoção predominante nos comentários foi a tristeza, com 20.8% do total, devido a influência dos acontecimentos do jogo, como a morte da personagem June. A emoção menos predominante foi o desgosto, com 3.5%, pois é a emoção que não tem muita ligação com o tema do jogo, justificando assim seu baixo percentual. No resultado dos sentimentos entre positivo e negativo, houve uma igualdade com 50% para cada um, explicado pela grande quantidade de comentários relacionados a emoções positivas, como a alegria, e a emoções negativas, como a tristeza. Em relação aos algoritmos de classificação, o que obteve melhor desempenho foi o *SVM* com o kernel radial, com o percentual de 79% de precisão, e o menos preciso foi o *SVM* com kernel *polynomial*, com o percentual de 34%. Os desempenhos dos algoritmos de classificação dependem do contexto em que são utilizados, fatores como tipos e volume dos dados podem influenciar, positivamente ou negativamente, nos resultados.

REFERÊNCIAS

[1] Sarmet, Mauricio Miranda and Pilati, Ronaldo. Efeito dos jogos digitais no comportamento: análise do General Learning Model. *Temas em Psicologia*, 24, 1 (mar 2016), 17-31.

[2] Krznaric, Roman. *Empatia: Sobre a arte de viver*. Zahar, Rio de Janeiro, 2015.

[3] Alencar, Vagner de. *Jogo ensina estudantes a se tornarem mais empáticos*. URL https://porvir.org/jogo-ensina-estudantes-se-tornarem-mais-empaticos/.

[4] Silva, Leandro Augusto da, Boscarioli, Clodis, and Peres, Sarajane Marques. *Introdução à Mineração de Dados: Com Aplicações em R*. Elsevier, Rio de Janeiro, 2017.

[5] Rezende, Solange Oliveira. *Sistemas inteligentes: fundamentos e aplicações*. Manole, Barueri, 2005.

[6] Behar, Patrícia Alejandra. *Modelos pedagógicos em educação a distância*. Artmed, Porto Alegre, 2009.

[7] Picard, Rosalind W. *Affective Computing*. The MIT Press, Cambridge, 2000.

[8] Cooper, Donald R. and Schindler, Pamela S. *Métodos de Pesquisa em Administração*. AMGH, Porto Alegre, 2016.

[9] Castro, Leandro Nunes de and Ferrari, Daniel Gomes. *Introdução à mineração de dados: conceitos básicos, algoritmos e aplicações*. Saraiva, São Paulo, 2016.

[10] Souza, Renato Rocha and Café, Lígia Maria Arruda. Análise de sentimento aplicada ao estudo de letras de música. *Informação & Sociedade: Estudos*, 28, 3 (Dezembro 2018), 275-286.

[11] Santos, Danilo Barros dos, Maciel, Cristiano, Pereira, Vinicius Carvalho, and Nunes, Eunice Pereira dos Santos. Analysis of the Perception of Users of Empathic Games in Discussion Forums and their Relation to Death. *7th Brazilian Symposium on Human Factors in Computing Systems (IHC 2018)* (October 22-26 2018), 9.

[12] Aguiar, Erikson Júlio de, Faiçal, Bruno S., Ueyama, Jó, Silva, Glauco Carlos, and Menolli, André. Análise de Sentimento em Redes Sociais para a Língua Portuguesa Utilizando Algoritmos de Classificação. *SIMPÓSIO BRASILEIRO DE REDES DE COMPUTADORES E SISTEMAS DISTRIBUÍDOS (SBRC)* (Maio 2018). Anais do XXXVI Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos.

[13] Rocha, Ana Carolina Pereira. *Mineração de textos para classificação de processos judiciais trabalhistas.* Dissertação (Mestrado Profissional em Computação Aplicada)—Universidade de Brasília, Brasília, 2019.

[14] Chicken, Rubber. *"Last Day of June": o amor não termina com a morte*. 2017.

[15] Canto, Carlos Eurico. *As linguagens mais populares em Ciência de Dados*. 2017.

[16] Amaral, Fernando. *Introdução à Ciência de Dados: Mineração de Dados e Big Data*. Alta Books, Rio de Janeiro, 2016b.

[17] Amaral, Fernando. *Aprenda Mineração de Dados: Teoria e prática.* Alta Books, Rio de Janeiro, 2016a.

[18] Santos, Tatiane Gomes Dos. *Análise de Opiniões Utilizando Técnicas de Mineração de Dados em Redes Sociais. Estudo de Caso: Twitter*. Anápolis, 2017.

[19] Silva, Rafael Medeiros Jacomel de Oliveira. *Análise de Sentimento em tweets*. Campinas, 2018.

[20] Agresti, Alan and Finlay, Barbara. *Métodos estatísticos para as Ciências Sociais*. Penso, Porto Alegre, 2012.

[21] Barros, Anna Carolina, Mattos, Daiane Marcolino de, Oliveira, Ingrid Christyne Luquett de, Ferreira, Pedro Guilherme Costa, and Duca, Victor Eduardo Leite de Almeida. *Análise de séries temporais em R: curso introdutório*. Elsevier, Rio de Janeiro, 2018.

[22] Bari, Anasse, Chaouchi, Mohamed, and Jung, Tommy. *Análise Preditiva Para Leigos*. Alta Books, Rio de Janeiro, 2019.

[23] Yates, Ricardo Baeza and Neto, Berthier Ribeiro. *Recuperação de Informação: Conceitos e Tecnologia das Máquinas de Busca*. Bookman, Porto Alegre, 2013.

[24] Pierson, Lillian. *Data science para leigos*. Alta Books, Rio de Janeiro, 2019.