

# Técnicas de Classificação para Análise de Sentimentos em Tweets em Português Brasileiro

Daniel Pimentel Kansaon<sup>1</sup>, Michele Amaral Brandão<sup>1</sup>

<sup>1</sup> Instituto de Ciências Exatas e Informática - Pontifícia Universidade Católica de Minas Gerais (PUC MG)

Av. Afonso Vaz de Melo, 1200 - Barreiro, Belo Horizonte – MG, 30640-070

daniel.kansaon@sga.pucminas.br

**Resumo.** Com a popularização das redes sociais online e o grande volume de dados produzidos a cada dia, a análise dessas informações pode ajudar a entender fenômenos, prever tendências e avaliar o senso comum. Na análise de sentimentos existem vários trabalhos para a língua inglesa, isso se dá devido a quantidade de ferramentas e técnicas existentes para esse idioma. No caso do português brasileiro, a quantidade de trabalhos é menor, devido não existirem tantas bases de dados à disposição. Muitas vezes as técnicas de classificação e análise de sentimentos apresentam resultados inferiores quando comparados a aplicação na língua inglesa. Neste trabalho, é realizado um estudo a fim de utilizar técnicas de análise de sentimentos e mineração de dados para descobrir sentimentos expressos diretamente ou indiretamente em tweets no idioma português brasileiro. Os tweets foram coletados através da API oficial do Twitter. Os dados foram armazenados em uma base de dados e organizados em classes de sentimentos para aplicação dos algoritmos de classificação. Foi necessário realizar processamentos nos tweets antes da aplicação dos algoritmos. Ademais, os emojis precisaram de um tratamento específico durante o trabalho. Os resultados mostram até 85% de acerto ao classificar sentimentos distintos. Por outro lado, as classes que compararam sentimentos próximos, não obtiveram resultados satisfatórios devido a semelhança dos sentimentos. Assim, a detecção de múltiplos sentimentos em textos ainda é um desafio da análise de sentimentos.

**Palavras-Chave:** Análise de sentimentos. Algoritmos de classificação. Classes de sentimentos. Descoberta de conhecimento. Mineração de dados. Mineração de opinião. Redes sociais. Twitter. Weka.

## 1. Introdução

Dados são coleções de valores em forma bruta, eles podem ser explorados e processados a fim de obter padrões, associações, mudanças e anomalias em prol de algum objetivo. Nas últimas décadas devido a globalização, a produção de dados na forma digital tem aumentado exponencialmente, esses dados vêm crescendo cada dia mais. Algumas estimativas apontam que a quantidade de informação produzida no mundo se duplica a cada 20 meses. Um bom processamento e manipulação dessas informações, pode trazer benefícios a organizações, grupos de investidores e até mesmo pessoas que desejam

abstrair algum tipo de informação sobre um conjunto de dados [DWIVEDI; KASLIWAL; SONI, 2016].

Segundo Dwivedi, Kasliwal e Soni (2016), a análise de dados é uma ciência que explora dados brutos e elicitando informações úteis e padrões ocultos. Ela permite analisar grandes volumes de dados que convencionalmente soluções de inteligência de negócios não podem fazer. A análise é usada em diferentes domínios como na ciência, negócios e ciências sociais. Com o avanço tecnológico e a popularização das redes sociais online, as pessoas utilizam cada vez mais essas redes como um grande aliado para a comunicação, fazendo com que as redes sejam uma grande fonte de informação.

Devido a sua grande quantidade de dados disponíveis, as análises dessas redes podem trazer informações valiosas para as várias empresas que buscam informações sobre seus produtos, clientes e que buscam entender melhor o mercado, os concorrentes, entre outros. Uma das técnicas que auxiliam essas avaliações é a análise de sentimentos que tem como ideia principal extrair e descobrir qual a opinião expressa em um texto. Assim, é possível classificar a sua polaridade [CAVALCANTI et al., 2012]. Essa técnica pode ser empregada em diversas situações, principalmente para entender o senso comum sobre um determinado acontecimento, como: eventos, economia, política.

Uma das técnicas que pode ser aplicada no contexto das redes sociais é a análise de sentimentos, que está associada diretamente com a mineração de dados. Segundo Calvalcanti et al (2012) a análise de sentimento é um estudo computacional que realiza uma análise de opiniões expressas em fontes não estruturadas. A ideia inicial desse tipo de análise não é simplesmente determinar qual tema principal ou enumerar palavras-chaves contidas em um texto. Isto pode ser feito utilizando técnicas como classificação de texto. A ideia principal é extrair e descobrir qual a opinião expressa no texto, assim podendo classificar a sua polaridade.

A análise de sentimento pode ser aplicada em vários contextos, um deles é nas redes sociais. A aplicação dessa análise pode ajudar avaliar tendências, opiniões sobre um determinado tema, além de fornecer uma visão geral sobre um determinado assunto. Nesse contexto, este trabalho apresenta uma mineração e análise de sentimentos de informações contidas em *tweets*, postagens realizadas por um usuário da rede social Twitter. Foram analisados sentimentos e opiniões expressos pelos usuários em um determinado contexto selecionado, encontrando nos *tweets* palavras que representam o respectivo sentimento analisado no contexto selecionado. A identificação dos sentimentos tem por apoio processos de mineração, que visam extrair informações relevantes, utilizando-se de métodos de classificação para determinar o sentimento expresso no *tweet*.

Nesse contexto, o foco deste trabalho é responder à pergunta de pesquisa: é possível descobrir sentimentos em *tweets* em português brasileiro utilizando técnicas de classificação? Em outras palavras, este trabalho realiza uma pesquisa exploratória cujo objetivo principal é verificar o desempenho de diferentes algoritmos de classificação quando aplicados na análise de sentimentos em *tweets* em português brasileiro.

O restante deste trabalho está organizado da seguinte forma. A seção 2 descreve conceitos importantes abordados no trabalho. A seção 3 descreve a metodologia utilizada, juntamente com a coleta e processamento dos dados. A seguir, a seção 4

descreve os resultados obtidos nos experimentos, detalhando a aplicação de cada método. A seção 5 discute os resultados experimentais obtidos. Finalmente, a seção 6 apresenta conclusões e comenta sobre trabalhos futuros.

## **2. Fundamentação Teórica**

Para melhor compreensão do trabalho, é importante o conhecimento de conceitos de mineração de dados e análise de sentimentos. Esta seção apresenta os principais conceitos abordados no trabalho, também apresenta uma contextualização da fonte de dados, que são as redes sociais.

### **2.1. Mineração de dados**

A mineração de dados é um processo de descoberta de padrões de dados. Esse processo deve ser automático ou semiautomático. A mineração de dados é a exploração e análise de grandes quantidades de dados a fim de descobrir padrões úteis e compreensíveis nos dados. Dentro da descoberta de conhecimento, o que é recuperado não é explícito no banco de dados, pelo contrário, são padrões implícitos. A mineração de dados encontra esses padrões e relacionamentos usando ferramentas de análise de dados e técnicas para construir modelos [HAMOUD; ATWELL, 2016].

Segundo Gulati (2016), a mineração de dados é um conceito amplo que consiste em série de passos. Inicialmente, os dados são pré-processados e depois várias técnicas de extração são aplicadas. Os resultados de técnicas de mineração são avaliados e interpretados.

O crescimento exponencial da geração de dados fez com que aumentasse o interesse na área de descoberta de conhecimento e mineração de dados. Existem várias técnicas e vários métodos para a descoberta de conhecimento, uma delas é a predição. Esse método pode ser usado em diversos problemas como por exemplo, na identificação do preço futuro de produtos, na previsão de dados financeiros. Ela é uma das atividades fundamentais em mineração de dados, juntamente com a programação genética (GP).

Um exemplo da aplicação da mineração de dados é em organizações e empresas, que mostram que a capacidade de extrair significado dos dados pode ser um recurso valioso, pois conhecer o comportamento dos clientes, coletar feedback para melhorar os serviços e produtos é essencial para o sucesso da organização. Embora as organizações podem realizar entrevistas ou distribuir questionários para coletar *feedbacks* dos clientes, isso pode se mostrar ineficiente, pois além de gerar esforço e gastar tempo, o *feedback* pode não representar uma indicação precisa dos clientes, pois nem sempre o questionário pode ser respondido com precisão ou até mesmo clientes podem não expressar bem suas opiniões. A análise de opiniões nas redes sociais pode apresentar muito mais resultados, pois a rede social é aberta e global, onde o usuário se sente livre para expressar opiniões [ALHUMOUD; ALBUHAIRI; ALTUWAIJRI, 2015].

### **2.2. Redes sociais online**

As redes sociais em geral, tem diversas características em comum, essas redes são usadas para trocas de informações, opiniões, entre outros. Segundo Benevenuto et al (2011), a rede social online pode ser definida como a união de pessoas que se interagem através de um meio de comunicação. Com essa definição, pode-se afirmar que as redes

sociais online, vão muito mais além das redes conhecidas, como: Twitter, facebook, linkedin.

As redes sociais online, permitem que as pessoas criem e disseminem conteúdos livremente, além de contribuir para a globalização. Estas redes podem ser usadas para estudos e pesquisas de várias áreas, principalmente da ciência computação. Temas como, aplicações distribuídas, tráfego de rede, mineração de dados, interação homem-computador, podem ser analisados através destas redes [BENEVENUTO; ALMEIDA; SILVA, 2011].

Uma rede social online bastante popular é o Twitter, que é um blog social fundado em 2006, que permite aos usuários se expressarem e compartilharem opiniões e sentimentos em público, isto acontece através de um texto estilo SMS, que se chama *tweet*. Toda essa atividade online acontece de modo livre e instantaneamente [ALHUMOUD; ALBUHAIRI; ALTUWAIJRI, 2015].

A utilização do Twitter e demais redes sociais não se remete apenas às atividades dos usuários. Atualmente várias empresas investem no marketing digital, com o objetivo de divulgar promoções e novos produtos. Além disso, as redes sociais são usadas para realizações de pesquisas, verificação da satisfação dos clientes através das análises de opiniões, entre outras. Com uma boa utilização, essas redes se tornam uma aliada dos negócios.

### **2.3. Métodos para análise de sentimentos em redes sociais**

Para realizar uma análise de sentimentos, existem diversos métodos e ferramentas que apresentam resultados diferentes. Em uma análise, vários métodos podem ser utilizados para obter um resultado consistente.

Nas redes sociais, muitas vezes são utilizados *emoticons* para expressão nos textos. Esses *emoticons*, na maioria das vezes são baseados em faces e expressam algum sentimento, uma análise baseada neles pode ser uma das maneiras mais simples de identificar um sentimento. Outro método é o LIWC, que é uma ferramenta que analisa textos e estima componentes que representam emoções baseadas em um dicionário que contém diversas palavras e sua categoria. Assim como esses, existe uma série de outros métodos, como, Happiness Index, SentiWordNet, SASA, PANAS-t, SenticNet e SentiStrength [ARAUJO; GONÇALVES; BENEVENUTO, 2015].

Uma outra forma de realizar a análise é através do cálculo da relevância. Nem todas as palavras no texto analisado possui a mesma importância, geralmente os termos mais frequentes possui um significado mais relevante. Existem várias formas para realizar o cálculo, uma delas é a frequência absoluta (TF), esse termo representa a quantidade de vezes que o termo aparece em um documento. Já a frequência relativa leva em consideração o tamanho do documento analisado e normaliza os pesos de cada palavra. Por fim, a frequência inversa de documentos (IDF), considera a quantidade de documentos onde o termo aparece. Com esse cálculo, considera-se como mais importante os elementos que aparecem em poucos documentos e como menos importante os elementos que mais aparecem nos documentos comparados. Com essas fórmulas é possível identificar as palavras mais relevantes do texto, assim é possível identificar o contexto e o sentimento expresso [MORAIS; AMBRÓSIO, 2007].

## 2.4. Ferramentas para a mineração de dados

Atualmente, existem vários tipos de ferramentas disponíveis para a realização da mineração dos dados. As ferramentas são usadas para manipular e encontrar informações relevantes a partir do conjunto de dados fornecido. Algumas ferramentas também podem gerar relatórios para resumir e proporcionar uma melhor visualização. Cada ferramenta possui características específicas, a escolha da ferramenta a ser usada dependerá do problema e das restrições. As principais ferramentas disponíveis para análise de dados são RapidMiner, WEKA, KNIME, R tool, Tableau public, OpenRefine, Solver, Orange, import.io, entre outras.

WEKA é uma plataforma desenvolvida pela Universidade de Waikato, que contém ferramentas de visualização e algoritmos para análise de dados e modelagem preditiva, juntamente com interface gráfica para o usuário, facilitando o acesso as funções. Além disso, WEKA fornece um conjunto de muitos algoritmos de aprendizado de máquina de mineração de dados, incluindo o pré-processamento em dados, classificação, clustering e associação. É recomendada para iniciantes, por ser uma ferramenta robusta, contendo vários recursos embutidos que não requerem programação ou codificação, facilitando a utilização. A ferramenta WEKA é compatível com os sistemas Windows, Mac OS X e Linux [DWIVEDI; KASLIWAL; SONI, 2016].

## 2.5. O estado da arte da análise de sentimentos

A mineração de dados juntamente com a análise de sentimentos tem ampla aplicabilidade em várias áreas do conhecimento, sendo elas: jogos, educação, política, sempre buscando informações significativas dentro do enorme volume de dados existentes. Assim, existem diversos trabalhos recentes que enfatizam o uso da mineração de dados juntamente com técnicas de classificação.

Por exemplo, Bouazizi e Ohtsuki (2016) propuseram uma abordagem para quantificação de sentimentos no Twitter. Este trabalho destacou a importância da detecção de sentimentos em um *tweet* e propôs uma maneira de extrair diferentes sentimentos existentes. Foram coletados cerca de 40.000 *tweets*, além de atribuir um sentimento para cada um, sendo eles: amor, felicidade, diversão, entusiasmo, alívio, ódio, raiva, tristeza, tédio preocupação, surpresa e neutralidade. No trabalho, foi utilizado a técnica de classificação Random Forest. Em primeiro passo, foi examinado o desempenho da classificação através de um conjunto de várias características. Foi observado que ao detectar sentimentos negativos, o método apresentou um desempenho superior aos outros sentimentos, mas de forma geral houve um desempenho bom na classificação de todos os sentimentos. Segundo os autores, a maior precisão da classificação do sentimento negativo se deu pela quantidade consideravelmente grande para os dados de treinamento dessa classe. Os sentimentos neutros foram difíceis de detectar e identificar, isso porque os sentimentos positivos e negativos podem ser considerados como dois polos opostos, enquanto a classe neutra está entre as duas.

Na pesquisa feita por Yadollahi, Shahraki e Zaiane (2017), foi sugerida uma categorização cuidadosa de tarefas na análise de sentimentos, também foi retratado a importância de uma taxonomia clara e lógica dentro dessa análise. Foram apresentados também métodos e melhorias da última geração para a análise de sentimento em textos. Nessa área existem duas principais subcategorias: a mineração de opinião e mineração

emocional. A primeira área aborda a expressão de opiniões, já a última, trata da articulação das emoções. A área de mineração emocional de texto ainda é bem recente e está sendo estudada cada vez mais. No trabalho, primeiramente foi explicado os elementos-chave da polaridade na classificação de sentimentos. Em seguida, foi apresentado um conjunto de recursos importantes, como: a base de dados que deve ser usada para uma boa análise e um dicionário léxico, que é um dicionário de palavras onde cada palavra está associada a uma pontuação que mostra seu grau de polaridade. Uma base de dados para classificação de texto, emoção e até mesmo identificação de polaridades, precisa estar bem representada contendo um número grande e bem representado de dados. Por fim, foram revisadas as teorias de emoções, que foram destacadas em duas teorias predominantes: a teoria da emoção discreta e a teoria do modelo dimensional. A teoria da emoção discreta afirma que diferentes emoções surgem de sistemas neurais separados, enquanto o modelo dimensional indica que um sistema neurofisiológico comum e interligado é responsável por todos os estados afetivos. Este modelo define emoções de acordo com uma ou mais dimensões em que geralmente uma delas se relaciona com intensidade das emoções.

No trabalho apresentado por Pandey e Rajpoot (2016), a análise de sentimentos foi feita através da comparação de diferentes técnicas de classificação. Os autores fizeram uma análise comparativa de vários algoritmos de classificação. Essa análise foi feita com a ajuda da ferramenta WEKA, utilizando um conjunto de dados que indicava o consumo de álcool por estudantes de uma escola. Os dados foram organizados em duas categorias, um conjunto de estudantes que consomem álcool e outro conjunto de estudantes que não consomem álcool. No trabalho foi usado vários algoritmos de classificação que já são incorporados no WEKA, como J48, Random Forest, Decision Tree, Random Tree, NaiveBayes, SimpleNaiveBayes, NaiveBayes, DecisionStump. A comparação dos algoritmos foi feita com a ajuda de vários parâmetros, como: precisão, estudantes classificados corretamente, estudantes classificados incorretamente. Por fim, o algoritmo que realizou a melhor classificação foi Decision Stump, com um resultado de 95,44% de acurácia.

Outro estudo que apresentou uma comparação de várias técnicas de classificação, foi o trabalho desenvolvido por Garg e Khurana (2014), que realizaram um experimento para avaliar o desempenho de vários métodos de classificação. O objetivo era utilizar algoritmos de classificação para projetar um modelo de detecção de intrusão eficaz, impedindo que as redes de computadores sejam invadidas, pois devido ao enorme tráfego de dados nas redes, os dados ficam vulneráveis a vários tipos de ataques. No trabalho foram avaliados os métodos de classificação que foram compatíveis com o conjunto de dados usados no trabalho. Os métodos foram avaliados através WEKA usando 41 atributos e cerca de 94.000 dados para um conjunto de treinamento, além de 48.000 dados para o conjunto de teste. Várias técnicas foram aplicadas para classificar diferentes métodos de acordo com vários parâmetros, como: Acurácia, ROC, Kappa, Tempo gasto para o treinamento dos dados, Erro absoluto, FPR e Recall, que são valores obtidos após a aplicação de cada método. O conjunto de dados utilizado para avaliação experimental consiste em instâncias selecionadas aleatoriamente do conjunto de dados NSL-KDD. Ao final do trabalho, foi apresentado uma lista de 45 métodos de classificação ordenados pelos seus respectivos resultados.

Os cinco algoritmos que apresentaram melhores resultados foram: Rotation Forest, Random Tree, Random Committee, Random Forest, IBK.

O trabalho proposto se diferencia em alguns pontos dos trabalhos apresentados nesta seção. Os dados usados é um fator muito importante, se diferenciando dos outros trabalhos por estarem no idioma português brasileiro. Outro ponto é a metodologia empregada, através da coleta e o processamento dos dados, trabalhando com a classificação de classes de sentimentos.

### 3. Metodologia

Nesta seção, é descrito os processos de coleta dos dados, o processamento dos dados e a preparação dos dados para a realização dos testes. O trabalho realizou uma análise a fim de identificar os sentimentos e emoções expressos em *tweets* através de métodos de classificação. Para atingir o objetivo, foi necessário realizar uma série de etapas visando processar e adaptar os dados com o formato compatível pela ferramenta.

A Figura 1 exemplifica as etapas realizadas na metodologia. Cada etapa é representada por um quadrado, que pode ser executada mais de uma vez.



Figura 1. Passos da metodologia. (Elaborado pelo autor)

#### 3.1. Seleção dos sentimentos

A definição das classes de sentimentos analisados durante a pesquisa, tem como base os diversos sentimentos que podem ser expressos através dos *tweets*. As expressões podem ser identificadas de várias formas, através de *emojis*, *emoticons*, *hashtag*.

Os sentimentos analisados no trabalho, tiveram por critério obter *tweets* que possam indicar sentimentos expressos pelos usuários, assim, foram selecionados 7 sentimentos a serem analisados. São eles: Triste, Chateado, Amor, Feliz, Raiva, Inveja, Ironia. Para obter os *tweets* com os respectivos sentimentos selecionados, o critério usado foi buscar por *tweets* que contém a *hashtag* com o nome do sentimento indicado, assim os *tweets* que possuem as *hashtags* (#Triste, #Chateado, #Feliz, #Amor, #Raiva, #Inveja, #Ironia) serão considerados no trabalho.

#### 3.2. Coleta dos dados

Após a seleção dos sentimentos na etapa anterior, foi necessário realizar a extração e o armazenamento dos dados para serem analisados. A API oficial do Twitter fornece vários recursos para a coleta de *tweets*, através dela, usuários que se cadastram como desenvolvedores conseguem coletar *tweets* a partir de parâmetros disponibilizados. A API possui algumas restrições para a coleta dos dados, mas se apresenta suficiente para realização do estudo. A API limita um número de 100 *tweets* como resposta para cada consulta realizada.

Como critério de busca e seleção dos *tweets*, foram definidas algumas condições para que o *tweet* faça parte dos dados. Com isso, é necessário que atenda as seguintes restrições:

- O texto precisa estar em português.
- Possuir uma das *hashtags* definidas: #Triste, #Chateado, #Feliz, #Amor, #Raiva, #Inveja, #Ironia.
- O *tweet* precisa ser recente.

Os demais parâmetros disponíveis na API foram ignorados, pois eram irrelevantes para o contexto do projeto.

A Figura 2 apresenta o funcionamento do fluxo de coleta dos *tweets*. Toda a comunicação com API foi feita através do protocolo HTTP usando métodos de GET. Todas as respostas retornadas pela API foram armazenadas na base de dados.



**Figura 2. Detalhamento do fluxo de coleta dos dados. (Elaborado pelo autor)**

Além da API oficial, existe o serviço de coleta dos dados e a base de dados. O Serviço de coleta dos dados foi desenvolvido para consultar a API oficial do Twitter, por onde obteve-se todos os *tweets* trabalhados no projeto. A comunicação do serviço com a API tem como parâmetro de busca vários critérios de seleção dos *tweets*, sendo eles: linguagem do *tweet*, *hashtag*, tipo do *tweet* entre outros. O serviço foi desenvolvido na linguagem C#, utilizando a biblioteca “*TweetSharp*” na versão 2.3.1 para facilitar as requisições feitas para a API. Após a requisição ser feita, a resposta obtida era armazenada na base de dados, assim como mostrado na Figura 2.

**Banco de dados:** A base de dados foi criada para armazenar todos *tweets* coletados durante o projeto. O banco de dados foi criado na linguagem SQL Server e foi hospedado no servidor na nuvem Azure. Todas as informações relevantes dos *tweets* retornados pela API, foram armazenadas na base de dados, sendo elas (texto, *hashtag* de busca, autor, data de criação, linguagem, quantidade de URLs, quantidade de menções feitas, texto das URLs, tipo).

A coleta foi feita entre os meses de agosto e outubro de 2017, contando com 12.814 *tweets* armazenados na base de dados. Essa quantidade de *tweets* inclui todos os sete sentimentos considerados na pesquisa.

### 3.3. Processamento dos dados

O processamento realizado tem como objetivo tratar os dados coletados removendo as informações que não contribuem para classificação dos dados. Cada *tweet* contém uma grande quantidade de informações que são irrelevantes para a análise. Por isso, são removidas dos *tweets*. Nesta etapa foram necessários vários tipos de processamentos e ajustes nos dados, conforme mostra a Figura 3 que demonstra todas as etapas realizadas de forma sequencial.

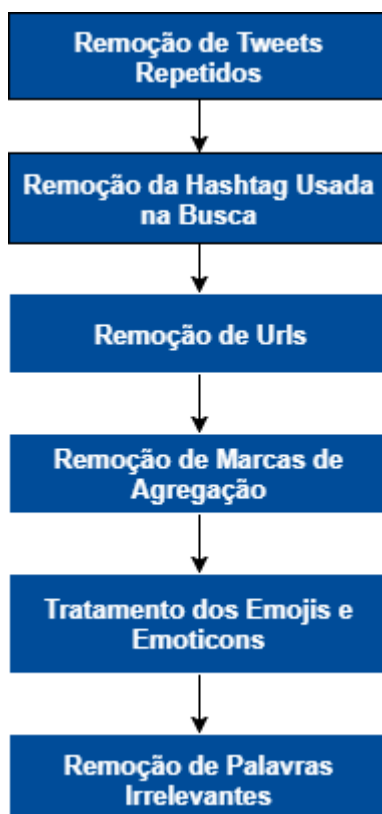


Figura 3. Etapas do processamento dos dados. (Elaborado pelo autor)

Todos os dados coletados foram submetidos as etapas descritas abaixo:





- Remoção de *tweets* repetidos: Ao coletar os *tweets*, foi observado que a API do Twitter retorna informações repetidas. Com isso, é necessário remover todos os *tweets* que se repetem na base de dados, para que não influencie o resultado final ao aplicar os métodos.
- Remoção da *hashtag* usada para busca: Como um dos critérios de busca foi pela *hashtag* do sentimento, foi necessária a remoção da *hashtag* no texto dos *tweets*.
- Remoção de URLs: Muitas vezes os usuários compartilham links ao postar um *tweet*, sendo assim, o link compartilhado fica visível no texto. Por ser uma informação desnecessária, os links foram removidos do texto de cada *tweet*.
- Remoção de marcas de agregação (RT, Mentions): Os usuários podem mencionar outros usuários no Twitter ou até mesmo replicar o *tweet* escrito por alguém. Essas duas ações são identificadas através da sigla RT ou @ nome do

usuário, contida no texto postado. Essas informações foram removidas, pois também eram desnecessárias para a análise.

- **Tratamentos dos *emojis* e *emoticons*:** Um sentimento pode ser expresso de várias formas em um *tweet*, uma delas é através de *emojis* ou *emoticons*, que são expressões usadas para se comunicar em um *tweet*. Os *emojis* são caracteres no formato Unicode, por isso, a ferramenta WEKA não consegue identificar esses caracteres. Assim, foi necessário processar e transformar os caracteres em palavras-chave, substituindo o *emoji* em uma palavra chave relacionada<sup>1</sup>.
- **Remoção de palavras irrelevantes:** Em um texto escrito, existem palavras que são irrelevantes para o conjunto de resultado esperado. Em *tweets* escritos em português, existem preposições como: a, as, de, que são irrelevantes para os métodos de classificação. Essas palavras foram removidas dos textos dos *tweets*.

Na Tabela 1 é mostrado como foi feita a transformação dos *emojis* em palavra-chave, exemplificando apenas alguns *emojis* que foram trabalhados, isso porque ao todo, foram encontrados 77 *emojis* que precisaram ser tratados. A primeira coluna da Tabela 1 mostra o *emoji* a ser tratado. Em seguida, a coluna “Palavra-Chave” representa a palavra que substituiu o respectivo *emoji* representado.

**Tabela 1. Processamento dos *emojis* em palavras-chaves.**

<b>Emoji</b>	<b>Palavra-Chave</b>
	ESorridente
	ECoracao
	EBravo
	EChorando

Após as etapas anteriores, com a ajuda da ferramenta WEKA, um filtro conhecido como “*StringToWordVector*” foi aplicado sobre os dados a serem executados pelos algoritmos. Quando esse filtro é aplicado sobre os dados, os atributos existentes são transformados em um vetor de valores numéricos, representando a ocorrência de cada palavra. Através desse filtro foi aplicado um outro filtro, conhecido como TF-IDF

---

<sup>1</sup> As informações sobre os *emojis* estão disponíveis no site <https://unicode.org/>, uma organização sem fins lucrativos que coordena o desenvolvimento do padrão Unicode.

que analisa a importância de cada palavra em um determinado texto [WITTEN et al., 2011].

Além das etapas de processamento, foi criado um critério para nivelamento dos dados, considerando a quantidade de *tweets* coletados durante o período de coleta. É necessário que os sentimentos tenham uma quantidade suficiente de dados para serem analisados, pois um número pequeno de dados pode influenciar em um resultado ruim. Para evitar essa situação, foi definida uma premissa. Para que um sentimento seja analisado, é necessário que possua um mínimo de 1000 *tweets*, para que tenha uma quantidade suficiente para treinar e testar os *tweets*. Com isso, os sentimentos: Ironia, Inveja e Raiva foram excluídos, portanto os métodos de classificação não serão aplicados nesses sentimentos, conforme mostra a Tabela 2.

Após as etapas de processamento, a Tabela 2 apresenta os dados armazenados e trabalhados no projeto. Após as etapas de processamento, a coluna “Quantidade Disponível Após Processamento” é obtida, onde contém a quantidade de dados que foram considerados no trabalho. Já a coluna “Quantidade Coletada” representa os dados brutos que foram coletados e armazenados na base de dados. Os valores que estão riscados significam que não foram considerados no projeto por causa da pouca quantidade de dados.

**Tabela 2. Detalhamento dos dados trabalhados no projeto.**

<b>Sentimento</b>	<b>Quantidade Coletada</b>	<b>Quantidade Disponível Após Processamento</b>
Chateado	2205	2050
Triste	3010	2800
Feliz	2135	2000
Amor	2975	2781
<b>Ironia</b>	<b>825</b>	<b>728</b>
<b>Raiva</b>	<b>732</b>	<b>521</b>
<b>Inveja</b>	<b>932</b>	<b>849</b>
Somatório	12.814	9631

### **3.4. Preparação dos dados**

Os dados coletados foram separados em classes de sentimentos. Alguns grupos de sentimentos foram criados para avaliar se os métodos de classificação conseguem identificar as diferenças entre os sentimentos. Os grupos foram formados visando comparar os sentimentos separando em: positivos x negativos, positivos x positivos e negativos x negativos. Dessa forma, os grupos formados para a análise são:

- Amor x Triste

- Feliz x Triste
- Feliz x Chateado
- Amor x Feliz
- Triste x Chateado

Após a seleção dos grupos de sentimentos a serem analisados, é necessário exportar os dados para o formato “.arff”, compatível com a ferramenta WEKA. Para isso, foi criado um arquivo para cada grupo de sentimento selecionado anteriormente, com os requisitos exigidos pelo WEKA. Neste arquivo, os dados foram divididos em dois atributos, o primeiro chamado “Descricao” onde contém o texto do *tweet* a ser analisado. O segundo atributo, chamado de “Sentimento” que define qual categoria de sentimento o *tweet* se enquadra.

A Figura 4, mostra o formato dos dados trabalhados após a inclusão no formato compatível pelo WEKA. Os atributos usados são declarados através da notação “@ATTRIBUTE”, já os dados, contém a notação “@DATA” que é declarada antes da primeira informação ser incluída.

```
@RELATION ClasseChateadoTriste

@ATTRIBUTE Descricao STRING
@ATTRIBUTE Sentimento {Chateado, Triste}

@DATA
"minha provas estão chegando ????",Chateado
```

Figura 4. Formato dos dados no padrão WEKA. (Elaborado pelo autor)

## 4. Resultados

Após a realização de todas as etapas de processamento dos dados, foi feita a aplicação dos algoritmos de classificação, juntamente com a análise dos resultados. Nesta seção, foi feita a avaliação dos algoritmos quanto a sua capacidade de classificar e diferenciar os sentimentos expressos nos *tweets* de forma correta.

### 4.1. Algoritmos aplicados e treinamento dos dados

A escolha dos algoritmos de classificação teve como base os algoritmos baseados no modelo Naive Bayes, que é uma técnica de classificação muito popular. Uma das suas principais características é a sua rapidez, precisão e fácil implementação. Esse algoritmo é um dos mais eficientes para problemas de probabilidade e aprendizagem de máquina. O algoritmo utiliza a probabilidade condicional para criar o modelo de dados a ser trabalhado. Inicialmente, o algoritmo trabalha com informações prévias, após obter novas informações, o teorema é aplicado obtendo os resultados. Uma das restrições desse algoritmo é que ele não leva em consideração o número de ocorrências de cada palavra, que é uma informação potencialmente útil ao determinar a categoria de um documento. Esse problema foi resolvido através da criação de outras técnicas que melhoram o desempenho desse modelo. Um exemplo de técnicas baseadas nesse

modelo de classificação é o Naive Bayes Multinomial, onde a classe de um documento é determinada não apenas pelas palavras existentes, mas também pelo número de vezes que ocorrem. Em geral, ele funciona melhor do que o modelo comum Naive Bayes [WITTEN et al., 2011].

Além da escolha dos algoritmos baseados no Naive Bayes, alguns algoritmos que apresentaram bons resultados em outros trabalhos, também foram selecionados. Isso permite verificar se os resultados obtidos nos trabalhos se repetem na classificação de *tweets* no idioma português brasileiro. No trabalho feito por Garg e Khurana (2014), são avaliados o desempenho de vários métodos de classificação, conforme descrito na Seção 2.5. Assim, foram escolhidos os algoritmos que apresentaram bons resultados e que são compatíveis com o formato dos dados trabalhados. Sendo eles: IBK, Random Forest, Random Committee. Existem várias categorias de classificadores, cada categoria contém uma aprendizagem de máquina diferente. O algoritmo IBK que foi selecionado, é encontrado na categoria Classificação Lazy. Esses tipos de classificadores são simples e eficazes, são usados principalmente para classificação de fluxos de dados. Já o algoritmo Random Committee, está na categoria Meta Classificador. Esses tipos de classificadores encontram o melhor conjunto de atributos para treinar, combinando as decisões de vários tipos diferentes métodos de classificação. Por fim, o Random Forest se encontra na categoria de Árvore de Decisão, que é uma técnica de classificação popular, na qual uma árvore é montada como resultado, cada nó denota um valor de um atributo e os ramos apresentam o resultado do teste [GARG; KHURANA, 2014].

Assim, os algoritmos selecionados para a classificação dos *tweets* foram: Naive Bayes, Naive Bayes Multinomial, Naive Bayes Multinomial Updateable, Sparge Generative Model, DMNB Text, Complement Naive Bayes, Bayesian Logistic Regression, IBK, Random Forest, Random Committee.

Após a seleção dos algoritmos é importante definir a forma que será feita os treinos dos *tweets*. Para treinar, é comum usar um terço dos dados para testes e dois terços para treinar, mas o problema dessa abordagem é que a parte usada para treino pode não ser representativa. O conjunto de dados deve ser representado na proporção certa, pois se o treinamento não estiver bem representado dificilmente o algoritmo apresentará um bom resultado. Uma maneira de mitigar qualquer parcialidade causada pela amostragem, é utilizar o tipo de treinamento conhecido como validação cruzada. Para treinar os *tweets* do projeto, os dados foram separados em dez partes. Nesse caso, os dados são divididos em dez partições iguais e cada uma delas é usada para testes e o restante é usado para treinamento. Nove décimos dos dados foram usados para treinamento e um décimo para teste. Esse processo é repetido dez vezes para que no final, cada instância seja usada uma vez para testar. Para obter a melhor estimativa de erro, é indicado usar dez partições, pois vários testes já foram realizados em outros estudos, que contaram com múltiplas bases de dados e técnicas de aprendizagem, mostrando que dez é o número certo de partições [WITTEN et al., 2011].

#### **4.2. Relevância das palavras em cada sentimento**

A Tabela 3 representa as dez principais palavras que foram relevantes para a análise de cada sentimento. As palavras representadas foram determinantes para a diferenciação dos sentimentos ao aplicar os métodos de classificação. Nota-se que algumas palavras se

repetem em mais de um sentimento, por exemplo os *emojis*: ERostoDesanimado, ERostoFranzido, que aparecem no sentimento Triste e Chateado, respectivamente isso acontece porque os dois sentimentos são relativamente parecidos, pois quando se está chateado o sentimento de tristeza também pode ser expresso, assim como os sentimentos Amor e Feliz. As palavras foram ordenadas pela relevância.

**Tabela 3. Palavras consideradas relevantes para a análise de cada sentimento.**

<b>Amor</b>	<b>Feliz</b>	<b>Triste</b>	<b>Chateado</b>
ECoracao ( <i>emoji</i> )	ESorridente ( <i>emoji</i> )	ERostoFranzido ( <i>emoji</i> )	ERostoDesanimado ( <i>emoji</i> )
ESorridente ( <i>emoji</i> )	ECoracao ( <i>emoji</i> )	EChorando ( <i>emoji</i> )	ERostoFranzido ( <i>emoji</i> )
Amor	Boas	ERostoDesanimado ( <i>emoji</i> )	Tristeza
#Love	Alegria	Triste	ESemExpressao ( <i>emoji</i> )
Amamos	Agradecer	Acabando	Poxa
#Paixao	Especial	Coitado	Triste
Amado	Feliz	ESemExpressao ( <i>emoji</i> )	Raiva
Desejo	Sorriso	Infelizmente	Ruim
Flores	Sucesso	Acabaram	Problemas
Fofura	Excelente	Aff	Sozinho

A Tabela 4 mostra a quantidade de *tweets* que foram considerados para cada classe de sentimento. A quantidade considerada foi padronizada, visando evitar a influência nos resultados pela desigualdade dos dados usados.

**Tabela 4. Quantidade de *tweets* para cada classe de sentimento.**

<b>Classe Sentimento</b>	<b>Quantidade Utilizada</b>
Amor x Triste	4000
Feliz x Triste	4000
Feliz x Chateado	4000
Amor x Feliz	4000
Triste x Chateado	4000

### 4.3. Aplicação dos algoritmos nas classes de sentimentos

As Tabelas 5, 6, 7, 8 e 9 apresentam os resultados obtidos através dos métodos de classificação, exibindo as variáveis resultantes para cada classe de sentimento. Nas Tabelas as colunas: Precision, F-Measure e ROC Area representam os resultados individuais para cada sentimento. Já a coluna Correct e Incorrect são valores para o algoritmo em geral. Os algoritmos que apresentaram melhores resultados estão destacados em negrito nas Tabelas.

**Tabela 5. Resultado da aplicação dos algoritmos na classe Amor x Triste.**

CLASSE – AMOR x TRISTE								
Algoritmo	Amor			Triste			Correct (%)	InCorrect (%)
	Precision	F-Measure	ROC Area	Precision	F-Measure	ROC Area		
Naive Bayes	0,86	0,59	0,87	0,65	0,77	0,87	70.72 %	29.27 %
Naive Bayes Multinomial	0,87	0,84	<b>0,92</b>	0,83	<b>0,86</b>	<b>0,92</b>	85.41 %	14.58 %
Naive Bayes Multinomial Updateable	0,88	0,84	<b>0,92</b>	0,83	<b>0,86</b>	<b>0,92</b>	85.54 %	14.45 %
Sparge Generative Model	0,84	0,82	<b>0,92</b>	0,83	0,84	<b>0,92</b>	83.84 %	16.15 %
DMNB Text	<b>0,89</b>	0,82	<b>0,92</b>	0,81	<b>0,86</b>	<b>0,92</b>	84.50 %	15.49 %
<b>Complement Naive Bayes</b>	0,85	<b>0,84</b>	0,85	0,85	<b>0,86</b>	0,85	<b>85.64 %</b>	<b>14.35 %</b>
Bayesian Logistic Regression	0,78	0,79	0,80	0,82	0,80	0,80	80.38 %	19.61 %
IBK	0,66	0,76	0,83	<b>0,87</b>	0,69	0,83	73.59 %	26.40 %
Random Forest	0,81	0,82	0,91	0,84	0,82	0,91	83.11 %	16.88 %
Random Committee	0,79	0,80	0,88	0,82	0,80	0,88	81.03 %	18.96 %

**Tabela 6. Resultado da aplicação dos algoritmos na classe Feliz x Triste.**

CLASSE – FELIZ x TRISTE								
Algoritmo	Feliz			Triste			Corre ct (%)	InCo rrect (%)
	Precisio n	F- Measure	ROC Area	Preci- sion	F- Measure	ROC Area		
Naive Bayes	0,76	0,54	0,79	0,70	0,79	0,79	72.10 %	27.89 %
Naive Bayes Multinomial	0,78	<b>0,74</b>	<b>0,87</b>	<b>0,82</b>	0,84	0,87	81.02 %	18.97 %
<b>Naive Bayes Multinomial Updateable</b>	0,80	<b>0,74</b>	<b>0,87</b>	0,81	<b>0,85</b>	<b>0,87</b>	<b>81.35</b> %	<b>18.64</b> %
Sparge Generative Model	0,75	0,72	0,86	0,81	0,83	0,86	79.31 %	20.68 %
DMNB Text	<b>0,85</b>	0,68	0,85	0,76	0,84	0,85	79.14 %	20.85 %
Complement NaiveBayes	0,75	0,75	0,79	0,83	<b>0,85</b>	0,79	80.34 %	19.65 %
Bayesian Logistic Regression	0,68	0,67	0,72	0,78	0,79	0,72	74.44 %	25.55 %
IBK	0,65	0,61	0,75	0,74	0,76	0,75	70.97 %	29.02 %
Ramdom Forest	0,76	0,69	0,84	0,78	0,82	0,84	77.96 %	22.03 %
Ramdom Committee	0,69	0,68	0,81	0,78	0,79	0,81	75.24 %	24.75 %

Tabela 7. Resultado da aplicação dos algoritmos na classe Feliz x Chateado.

CLASSE – FELIZ x CHATEADO								
Algoritmo	Feliz			Chateado			Corre ct (%)	InCo rrect (%)
	Preci- sion	F- Measure	ROC Area	Preci- sion	F- Measure	ROC Area		
Naive Bayes	0,83	0,59	0,79	0,60	0,72	0,79	67.09 %	32.90 %
<b>Naive Bayes Multinomial</b>	<b>0,81</b>	<b>0,80</b>	<b>0,87</b>	<b>0,77</b>	<b>0,79</b>	<b>0,87</b>	<b>79.60 %</b>	<b>20.39 %</b>
<b>Naive Bayes Multinomial Updateable</b>	<b>0,81</b>	<b>0,80</b>	<b>0,87</b>	<b>0,77</b>	<b>0,79</b>	<b>0,87</b>	<b>79.60 %</b>	<b>20.39 %</b>
Sparge Generative Model	0,79	0,77	0,86	0,74	0,77	0,86	77.17 %	22.82 %
DMNB Text	0,79	0,77	0,85	0,74	0,76	0,85	76.86 %	23.10 %
Complement NaiveBayes	0,81	0,80	0,79	<b>0,77</b>	<b>0,79</b>	<b>0,96</b>	79.54 %	20.45 %
Bayesian Logistic Regression	0,76	0,77	0,75	0,75	0,74	0,75	76.00 %	23.99 %
IBK	0,60	0,71	0,72	0,73	0,50	0,72	63.86 %	36.13 %
Ramdom Forest	0,73	0,76	0,83	0,75	0,72	0,83	74.58 %	25.41 %
Ramdom Committee	0,72	0,75	0,80	0,73	0,70	0,80	73.13 %	26.86 %

Tabela 8. Resultado da aplicação dos algoritmos na classe Amor x Feliz.

CLASSE – AMOR x FELIZ								
Algoritmo	Amor			Feliz			Corre ct (%)	InCo rrect (%)
	Preci- sion	F- Measure	ROC Area	Preci- sion	F- Measure	ROC Area		
Naive Bayes	0,69	0,74	0,72	0,66	0,58	0,72	68.23 %	31.76 %
Naive Bayes Multinomial	0,76	0,76	<b>0,80</b>	0,68	<b>0,69</b>	<b>0,80</b>	73.25 %	26.74 %
<b>Naive Bayes Multinomial Updateable</b>	0,76	<b>0,77</b>	<b>0,80</b>	0,69	<b>0,69</b>	<b>0,80</b>	<b>73.59</b> %	<b>26.40</b> %
Sparge Generative Model	<b>0,77</b>	0,73	0,78	0,64	0,68	0,78	71.41 %	28.58 %
DMNBText	0,71	0,76	0,78	<b>0,71</b>	0,63	0,78	71.46 %	28.53 %
Complement NaiveBayes	0,77	0,75	0,72	0,67	<b>0,69</b>	0,72	72.95 %	27.04 %
Bayesian Logistic Regression	0,72	0,73	0,68	0,65	0,63	0,68	69.50 %	30.49 %
IBK	0,65	0,73	0,70	0,66	0,51	0,70	65.90 %	34.09 %
Ramdom Forest	0,73	0,75	0,76	0,67	0,64	0,76	70.71 %	29.28 %
Ramdom Committee	0,71	0,73	0,74	0,65	0,62	0,74	69.15 %	30.84 %

**Tabela 9. Resultado da aplicação dos algoritmos na classe Chateado x Triste.**

CLASSE – CHATEADO x TRISTE								
Algoritmo	Chateado			Triste			Corre ct (%)	InCo rrect (%)
	Preci- sion	F- Measure	ROC Area	Preci- sion	F- Measure	ROC Area		
Naive Bayes	0,43	0,05	0,57	0,63	0,77	0,57	62.99 %	37.00 %
Naive Bayes Multinomial	0,52	<b>0,52</b>	<b>0,68</b>	<b>0,72</b>	0,72	<b>0,68</b>	65.16 %	34.83 %
<b>Naive Bayes Multinomial Updateable</b>	<b>0,54</b>	0,51	<b>0,68</b>	<b>0,72</b>	0,74	<b>0,68</b>	<b>66.48</b> %	<b>33.51</b> %
Sparge Generative Model	0,51	0,49	0,50	0,70	0,72	0,65	64.39 %	35.60 %
DMNB Text	0,57	0,29	0,63	0,66	0,77	0,63	65.39 %	34.60 %
Complement NaiveBayes	0,50	0,54	0,62	0,73	0,70	0,62	64.14 %	35.85 %
Bayesian Logistic Regression	0,48	0,47	0,58	0,69	0,70	0,58	62.13 %	37.86 %
IBK	0,41	0,31	0,53	0,64	0,71	0,53	59.58 %	40.41 %
Ramdom Forest	0,49	0,34	0,60	0,66	<b>0,83</b>	0,60	63.13 %	36.86 %
Ramdom Committee	0,47	0,32	0,59	0,65	0,80	0,59	61.12 %	34.87 %

Com base na aplicação dos algoritmos selecionados e da divisão das diversas classes de sentimentos, alguns algoritmos se destacaram na classificação, já outros não apresentaram bons resultados. Os algoritmos baseados no modelo de bayes, foram os

que apresentaram os melhores resultados, sendo eles: Naive Bayes Multinomial Updateable, Naive Bayes Multinomial e Complement Naive Bayes. A Tabela 10 mostra a média geral desses algoritmos levando em consideração a propriedade “classificação correta” de todas as classes montadas no projeto.

**Tabela 10. Média geral dos algoritmos que obtiveram melhores resultados.**

Algoritmo	Média geral
Naive Bayes Multinomial Updateable	77,31 %
Naive Bayes Multinomial	76,88 %
Complement Naive Bayes	76,52 %

As classes que compararam sentimentos distintos positivos x negativos, foram as que obtiveram melhores resultados. Em primeiro lugar, a classe amor x triste obteve o melhor resultado, chegando a 85.64% de acerto na classificação, com o método Complement Naive Bayes. Uma das razões para esse resultado, é a quantidade de *emojis* encontrados nas classes. Os *emojis* Ecoracao e Esorridente foram encontrados diversas vezes no sentimento amor, já no sentimento triste, não foi detectado nenhuma ocorrência desses *emojis*. As palavras relevantes para a classificação não se repetiram entre os sentimentos, conforme mostra a Tabela 3. Logo em seguida, a classe feliz x triste obteve 81.35% de acerto, com o método Naive Bayes Multinomial Updateable. Os *emojis* ECoração e ESorridente encontrados na classe feliz, também não são detectados na classe triste. Por fim, a classe feliz x chateado ficou em terceiro lugar, com 79.60% de acerto na classificação, com os algoritmos Naive Bayes Multinomial e Naive Bayes Multinomial Updateable. Vale ressaltar que as classes que continham os sentimentos positivos x negativos, as palavras relevantes para a classificação não se repetiram, logo, essas classes foram as que apresentaram os melhores resultados comparado com as demais classes. A Tabela 11, exibe os três algoritmos que apresentaram os melhores resultados, seguindo com a médias de acerto dos algoritmos nas classes que compararam sentimentos positivos x negativos.

**Tabela 11. Média das classes que compararam sentimentos positivos x negativos.**

Algoritmo	Média dos sentimentos diferentes
Naive Bayes Multinomial Updateable	82,16 %
Naive Bayes Multinomial	82,01 %
Complement Naive Bayes	81,84 %

Em contrapartida, as classes que compararam sentimentos parecidos positivos x positivos ou negativos x negativos obtiveram os piores resultados. Os algoritmos usados para a classificação não conseguiram distinguir ou encontrar alguma palavra que expressasse um sentimento de maneira que diferenciasse um do outro, fazendo com que os resultados das classes representadas por sentimentos semelhantes sejam inferiores as classes de sentimentos distintos. Nos sentimentos triste e chateado, quatro das dez palavras consideradas relevantes para a classificação se repetiram, sendo elas:

ERostoFranzido, ESemExpressao, Triste, ERostoDesanimado. Isso indica a semelhança entre os sentimentos, refletindo nos resultados obtidos. O método Naive Bayes Multinomial Updateable obteve o melhor resultado dos algoritmos aplicados na classe chateado x triste, com 66.48% de classificação correta. Já nos sentimentos feliz e amor os *emojis*: ECoracao e ESorridente apareceram em ambos sentimentos, indicando a relação entre eles. Apesar da semelhança, a classe amor x feliz obteve um resultado de 73.59% de classificação correta com o método Naive Bayes Multinomial Updateable, sendo superior ao resultado obtido pela classe triste x chateado. A Tabela 12 mostra a média dos três algoritmos que apresentaram os melhores resultados nas classes de sentimentos semelhantes.

**Tabela 12. Média das classes que compararam sentimentos negativos x negativos e positivos x positivos.**

Algoritmo	Média dos sentimentos semelhantes
Naive Bayes Multinomial Updateable	70,03 %
Naive Bayes Multinomial	69,20 %
Complement Naive Bayes	68,54 %

Os algoritmos IBK, Random Forest, Random Committee, não apresentaram resultados eficientes quando comparados com os resultados apresentados no trabalho feito por Garg e Khurana (2014). Isso se dá pelo fato do formato dos dados ser diferente, além de que, os dados usados estão em outro idioma. Outro ponto importante que distingue os dois trabalhos, é a quantidade de dados usados para as etapas de processamento e teste. No trabalho feito por Garg e Khurana (2014), foram considerados cerca de 142.000 para o conjunto de treino e teste.

Além das etapas de treino e de processamento dos dados, que foram importantes para a classificação dos sentimentos, os *emojis* se mostram essenciais para a classificação. Os *emojis* representaram as principais palavras por caracterizar um sentimento, diferenciando um sentimento positivo de outro negativo. Apesar dos métodos existentes produzirem resultados inferiores quando aplicados no idioma português brasileiro e da quantidade restrita dos dados coletados, os resultados foram satisfatórios, não sendo superiores a outras pesquisas, mas de acordo com outras pesquisas realizadas na área.

## 5. Conclusão

Neste trabalho, foram apresentados conceitos de mineração de dados e análise de sentimentos, que teve como objetivo analisar e classificar sentimentos expressos no Twitter. Técnicas de mineração de dados e classificação de texto foram aplicadas nos dados coletados. No trabalho, foram utilizados métodos de classificação baseados no modelo Naive Bayes e em outros modelos de classificação, como árvores de decisão, para classificar *tweets* em português brasileiro. Apesar de ser um tema extremamente popular, onde existem diversos estudos, ainda existem vários desafios para a descoberta de sentimentos.

O estudo contou com sete sentimentos, sendo eles: triste, chateado, amor, feliz, raiva, inveja, ironia. Os dados foram obtidos através do Twitter, onde foram coletados 12.814 *tweets* para o estudo. Ao final do processamento, os sentimentos inveja, ironia, raiva, foram excluídos, restando 9.631 *tweets*. Após a coleta, os dados foram separados em grupos de sentimentos: positivo x negativo, negativo x negativo e positivo x positivo, para assim, serem aplicados nos algoritmos de classificação. Os resultados obtidos foram satisfatórios, mostrando que os algoritmos baseados no modelo Naive Bayes apresentaram os melhores resultados. As classes que possuíram os sentimentos positivos x negativos foram as classes que obtiveram os melhores resultados, chegando a 85% de acerto na classificação com o método Complement Naive Bayes. Apesar dos resultados não serem superior a outras pesquisas, este estudo se mostra relevante, pois os dados usados são exclusivamente na língua portuguesa. Existem vários trabalhos na língua inglesa, mas no caso do português brasileiro, a quantidade de trabalhos é menor, devido não existirem tantas bases de dados à disposição [NERI et al. 2012].

Para os trabalhos futuros, o tema apresentado pode ser estudado explorando a coleta dos dados juntamente com a aplicação dos métodos de classificação, comparando com resultados obtidos em textos na língua inglesa. Métodos de análise de sentimentos podem ser utilizados a fim de realizar uma análise mais detalhada, estudando a importância dos *emojis e emoticons* na identificação dos sentimentos expressos nas redes sociais e a detecção de múltiplos sentimentos em *tweets*. A diferenciação das polaridades dos sentimentos, também pode ser uma análise relevante para o trabalho.

## 7. Referências

- ALHUMOUD, Sarah; ALBUHAIRI, Tarfa; ALTUWAIJRI, Mawaheb. Arabic sentiment analysis using WEKA a hybrid learning approach. In Procs. of International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K), 7., 2015, Lisboa. Anais... Lisboa: Ieee, 2015. p. 402 - 408.
- ARAÚJO, Matheus; GONÇALVES, Pollyanna; BENEVENUTO, Fabrício. Métodos para Análise de Sentimentos em Mídias Sociais. In Procs. of: Simpósio Brasileiro de Multimídia e Web (WEBMEDIA), 21, Manaus, 2015. p. 27 - 30.
- BOUAZIZI, Mondher; OHTSUKI, Tomoaki. Sentiment Analysis in Twitter: From Classification to Quantification of Sentiments within Tweets. In Procs. of Global Communications Conference (GLOBECOM), 2016 IEEE, 2016, Washington. Proceedings... Washington: Ieee, 2016. p. 1 - 6.
- BENEVENUTO, Fabrício; ALMEIDA, Jussara M.; SILVA, Altigran S. da. Coleta e Análise de Grandes Bases de Dados de Redes Sociais Online. In Procs. of Congresso da Sociedade Brasileira de Computação (CSBC), 2011, Natal. Jornada de Atualizações em Informática (JAI). Natal: Csbcc, 2011. p. 1 - 47.
- CAVALCANTI, Diana C. et al. Análise de Sentimento em Citações Científicas para Definição de Fatores de Impacto Positivo. In Procs. of IV International Workshop on Web and Text Intelligence, 4. 2012. p. 1 - 10.
- DWIVEDI, Shraddha; KASLIWAL, Paridhi; SONI, Suryakant. Comprehensive study of data analytics tools (RapidMiner, Weka, R tool, Knime). 2016 Symposium On

- Colossal Data Analysis And Networking (cdan), [s.l.], p.1-8, mar. 2016. Institute of Electrical and Electronics Engineers (IEEE). <http://dx.doi.org/10.1109/cdan.2016.7570894>.
- GARG, Tanya; KHURANA, Surinder Singh. Comparison of classification techniques for intrusion detection dataset using WEKA. In Procs. of International Conference on Recent Advances and Innovations in Engineering (ICRAIE-2014), 2014, Jaipur. Proceedings..Jaipur: Ieee, 2014. p. 1 - 5.
- GULATI, Hina. Predictive Analytics Using Data Mining Technique. International In Procs. of Conference On Computing for Sustainable Global Development. p.713-716, 2015. Institute of Electrical and Electronics Engineers (IEEE). <http://ieeexplore.ieee.org/document/7100342>.
- HAMOUD, Bothaina; ATWELL, Eric. Quran question and answer corpus for data mining with WEKA. 2016 Conference Of Basic Sciences And Engineering Studies (sgcac), [s.l.], p.221-226, fev. 2016. Institute of Electrical and Electronics Engineers (IEEE). <http://dx.doi.org/10.1109/sgcac.2016.7458032>.
- MORAIS, Edison Andrade Martins; AMBRÓSIO, Ana Paula L.. Mineração de Textos. 2007. 30 f. Dissertação (Mestrado) - Curso de Informática, Universidade Federal de Goiás, Goiânia, 2007.
- NERI, F; ALIPRANDI, C; CAPECI, F; CUADROS, M. Sentiment Analysis on Social Media, 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining.
- PANDEY, Anand Kishor; RAJPOOT, Dharmveer Singh. A comparative study of classification techniques by utilizing WEKA. In: SIGNAL PROCESSING AND COMMUNICATION (ICSC), 2016 INTERNATIONAL CONFERENCE ON, 2016, Noida. Proceedings... Noida: Ieee, 2016. p. 219 - 224.
- WITTEN I. H; FRANK, E; HALL, M. A. (2011). Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann Publishers, 3 Edition. ISBN: 978-0-12-374856-0.
- YADOLLAHI, Ali; SHAHRAKI, Ameneh Gholipour; ZAIANE, Osmar R.. Current State of Text Sentiment Analysis from Opinion to Emotion Mining. Acm Computing Surveys, [s.l.], v. 50, n. 2, p.1-33, 25 maio 2017. Association for Computing Machinery (ACM). <http://dx.doi.org/10.1145/3057270>.