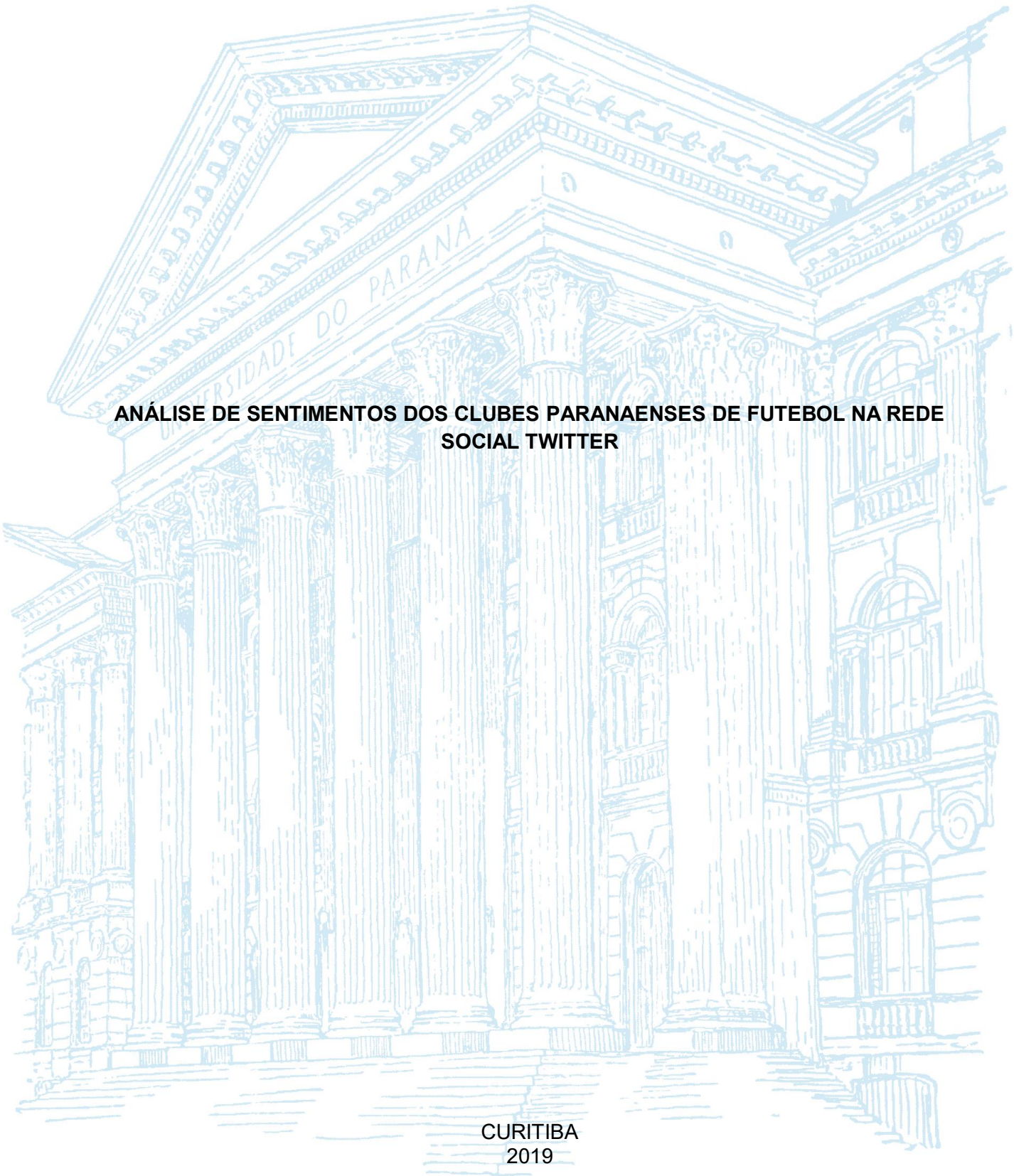
UNIVERSIDADE FEDERAL DO PARANÁ

GABRIEL LUIS DEGANI




# ANÁLISE DE SENTIMENTOS DOS CLUBES PARANAENSES DE FUTEBOL NA REDE SOCIAL TWITTER

CURITIBA
2019

GABRIEL LUIS DEGANI

# ANÁLISE DE SENTIMENTOS DOS CLUBES PARANAENSES DE FUTEBOL NA REDE SOCIAL TWITTER

Trabalho de Conclusão de Curso apresentado como requisito parcial para a obtenção do grau de bacharel no Curso de Gestão da Informação, do Departamento de Ciência e Gestão da Informação, do Setor de Ciências Sociais Aplicadas, da Universidade Federal do Paraná.

Orientadora: Prof.ª Dr.ª Denise Fukumi Tsunoda

CURITIBA
2019

## AGRADECIMENTOS

Primeiramente, agradeço à Universidade Federal do Paraná pela oportunidade e conhecimento obtido durante a graduação. Além dos ensinamentos com o curso, a Universidade me formou como cidadão e espero um dia ser capaz de retornar para a sociedade tudo que pude absorver durante meu período como estudante.

Agradeço à minha mãe por todo o apoio e incentivo durante minha trajetória na faculdade, tenho muito orgulho das suas conquistas e de ser seu filho.

Aos amigos pelos momentos de descontração e amizade, tornando os dias na Universidade mais leves e divertidos.

À professora Denise por todos os ensinamentos, orientações e dedicação ao longo do curso, principalmente durante as disciplinas de Trabalho de Conclusão de Curso I e II. Sentirei falta das conversas de manhã. Muito obrigado por tudo.

Por fim, aos colegas Josiel de Oliveira e Bruno Ábia por todo o auxílio durante as etapas desta pesquisa, sem vocês este trabalho seria ainda mais árduo.

## RESUMO

Apresenta um estudo exploratório visando identificar os sentimentos dos torcedores dos três principais clubes do estado do Paraná. Analisa as postagens feitas na rede social Twitter a fim de classificar as publicações como positivas, negativas ou neutras. Investiga a utilização de técnicas de análise de sentimentos dentro de instituições esportivas de maneira estratégica. Desenvolve um fluxo elucidando as etapas de criação das bases de dados, pré-processamento, processamento e pós-processamento dos dados coletados. Utiliza ferramentas como o Microsoft Excel® e a linguagem de programação Python para as etapas de pré-processamento e processamento, além do Orange Canvas e Power BI na etapa de pós-processamento. Aplica os algoritmos de aprendizado de máquina *Naïve Bayes*, SVM e CART, apresentando resultados satisfatórios na classificação das bases de dados, com taxas de acerto na classificação das instâncias superior a 65% para os três algoritmos utilizados. Evidencia a possibilidade do uso da mineração de opiniões como recurso para analisar o entusiasmo dos torcedores. Recomenda trabalhos futuros de análise de sentimentos em clubes de futebol com o fluxo proposto nesta pesquisa na exploração de outras redes sociais e coleta de maior quantidade de dados.



Palavras-chave: Análise de sentimentos. Twitter. Clubes de futebol. Aprendizado de máquina.

## ABSTRACT

It presents an exploratory study on sentiment analysis of fans of the three main clubs of the state of Paraná. It analyzes posts made on the social network Twitter in order to classify the posts as positive, negative or neutral. Look into the use of sentiment analysis techniques within sports institutions in a strategic way. It proposes to develop a flow by elucidating steps of database creation, preprocessing, processing and post processing of the collected data. It uses tools like Microsoft Excel® and Python programming language for the preprocessing and processing steps, as well as Orange Canvas and Power BI in the post processing steps. Applies Naive Bayes, SVM and CART machine learning algorithms, presenting satisfactory results in classification of databases, with rates higher than 65% for all the algorithms used. Evidence of using opinion mining as a resource to analyze the enthusiasm of fans. It recommends future sentiment analysis work in football clubs using the flow proposed in this research exploring other social networks and collecting more data to performing the analysis.



Keywords: Sentiment Analysis. Twitter. Soccer teams. Machine Learning.

## LISTA DE FIGURAS

## LISTA DE QUADROS

## LISTA DE SIGLAS

AM – Aprendizado de Máquina

*API – Application Programming Interface*

BOW – *Bag of Words*

CSV – *Common Separated Values*

FN – *False Negatives*

FP – *False Positives*

IBGE – Instituto Brasileiro de Geografia e Estatística

IDF – *Inverse Document Frequency*

NLTK – *Natural Language ToolKit*

PLN – Processamento de Linguagem Natural

POS – *Part of Speech*

SAC – Serviço de Atendimento ao Consumidor

SVM – *Support Vector Machine*

TF – *Term Frequency*

TF/IDF – *Term Frequency – Inverse Document Frequency*

TN – *True Negatives*

TP – *True Positives*

TSA – *Twitter Sentiment Analysis*

UFPR – Universidade Federal do Paraná

WWW- *World Wide Web*

## SUMÁRIO

# 1 INTRODUÇÃO

As redes sociais cresceram exponencialmente em popularidade ao longo dos anos, impulsionadas pelo aumento do poder aquisitivo da população mundial e pela melhoria e redução dos custos da infraestrutura de tecnologia e comunicação (CASTRO; FERRARI, 2016).

Dentre as diversas redes sociais disponíveis, o Twitter se destaca como uma das mais populares. O serviço de microblog permite ao usuário escrever sobre qualquer assunto dentro de 280 caracteres, além da capacidade de enviar fotos, vídeos e compartilhar a publicação de outros usuários. Somente no ano de 2014 as pessoas enviaram mais de 500 milhões de tweets (nome dados às postagens realizadas dentro da rede social) por dia (TWITTER, 2014). Dessa forma, os usuários das redes sociais deixaram de ser somente consumidores da informação para tornarem-se geradores de conteúdos através de suas opiniões sobre os mais diversos assuntos.

Castro e Ferrari (2016, p. 19) destacam que as opiniões dos usuários sobre os mais diversos assuntos representam uma grande utilidade para ramos como inteligência de marketing, comércio eletrônico e monitoramento de reputação, de maneira a qual a análise de dados no Twitter pode evidenciar o motivo da repercussão de determinados assuntos. Os autores afirmam que a aplicação de técnicas de mineração de dados permite extrair informações úteis e indispensáveis para a tomada de decisão estratégica.

Neste contexto, surge a mineração de textos, a qual fornece um conjunto de técnicas capazes de automatizar o processo de coleta e estruturação de informações, permitindo que as organizações possam saber o que os usuários estão comentando sobre elas em seus perfis na web (FILHO CARVALHO, 2014). Tal verificação contribui para a tomada de decisão das organizações sobre diferentes aspectos, como estratégias de marketing, fornecimento de serviços, exploração de mercado, dentre outros.

No caso dos clubes de futebol, estar ciente das opiniões nas redes sociais permite a adoção e a verificação de estratégias. Um clube pode desejar saber se alguma contratação está sendo aprovada pela torcida ou qual a opinião referente ao desempenho do time em determinada partida. Com essas informações recuperadas

através das redes sociais, o clube pode focar sua venda de materiais relacionados a um determinado jogador e realizar ações de marketing especificas.

Assim sendo, a presente pesquisa visa recuperar dados do Twitter alusivos aos clubes Athlético Paranaense, Coritiba e Paraná, buscando classificar os dados recuperados de acordo com seu sentimento (positivo, negativo ou neutro), a fim de identificar padrões, tendências e obter informações capazes de gerar tomadas de decisões.

## 1.1 PROBLEMATIZAÇÃO

Clubes de futebol são instituições que buscam gerar cada vez mais valor, sendo tal valor coletado por meio de patrocínios, vendas de produtos, transmissões televisivas, com seus torcedores e associados. Em 2016, os 20 principais clubes brasileiros obtiveram receita bruta total de R$ 4,86 bilhões (GONÇALVES, 2016). De acordo com a Revista Forbes (2018) em uma pesquisa realizada no ano de 2018 referente aos 50 clubes mais valiosos da américa, 15 representantes eram brasileiros, demonstrando a receita gerada pelas instituições no Brasil. Destes 15 representantes, 3 clubes estavam entre os 10 primeiros, corroborando na demonstração da receita gerada por essas organizações.

Devido ao tamanho dos clubes e sua capacidade de receita, seu interesse em questões relacionadas com marketing, imagem e mercado são de suma importância. Além disso, de acordo com uma pesquisa realizada em janeiro de 2012 pelo grupo Pluri Pesquisas Esportivas (2019), o Brasil possuía uma população de 192.7 milhões de habitantes no ano de 2012, dentre os quais, 151.8 milhões de pessoas identificam-se com algum clube de futebol, distribuídos entre as diversas equipes espalhadas pelo Brasil. Este número demonstra a popularidade do esporte no país, o qual exerce grande atração e envolvimento na população brasileira.

Na cidade de Curitiba, capital do estado do Paraná, de acordo com Rudnick (2017), 31,2% dos torcedores torcem para o clube Athlético Paranaense enquanto o rival Coritiba possui a segunda maior proporção com 26,7% e o Paraná Clube com 9%, sendo os clubes com as maiores torcidas da capital e do estado.

A recuperação de informações em redes sociais e sua consequente análise pode contribuir para averiguar todas as questões anteriormente elencadas, de modo que este projeto de pesquisa busca responder as seguintes questões: **pode-se**

**identificar os sentimentos dos torcedores dos principais clubes de futebol do Estado do Paraná por meio da análise de sentimentos nas redes sociais?**

## 1.2 OBJETIVOS

Para que a indagação proposta no problema de pesquisa possa ser respondida, faz-se necessário a definição dos objetivos, de modo a dar um direcionamento para a investigação. Dessa forma, são definidos o objetivo geral e os específicos, sendo que estes últimos são os passos para que o geral possa ser alcançado.

### 1.2.1 Objetivo Geral

O objetivo da pesquisa é identificar e analisar os sentimentos dos torcedores dos três principais clubes de futebol do Estado do Paraná em suas publicações no Twitter no período do pós-jogo.

### 1.2.2 Objetivos Específicos

Derivados do objetivo geral, foram definidos os específicos:

a) coletar dados no Twitter referentes aos clubes selecionados;
b) desenvolver um fluxo para coletar, tratar e processar os dados coletados do Twitter;
c) aplicar algoritmos de aprendizado de máquina, propondo modelos preditivos para classificar os tweets de acordo com o sentimento contido neles: negativo, neutro ou positivo.

## 1.3 JUSTIFICATIVA

A escolha do tema deu-se devido ao volume de dados e informações relevantes disponíveis nas redes sociais. Analisar sentimentos em redes sociais é um tema em desenvolvimento e que vem sendo cada vez mais explorada por diversas áreas de interesse, sendo um tema atual no contexto acadêmico e de negócios. Realizar explorações através da opinião de outras pessoas contribui para saber como o assunto é visto e reagido de modo geral.

Para selecionar os clubes analisados por este projeto de pesquisa, buscou verificar quais clubes possuíam as maiores torcidas no estado do Paraná através de pesquisas realizadas por institutos como o Instituto Brasileiro de Geografia e Estatística (IBGE) e o Paraná Pesquisas, além de sites especializados no tema. Ao final da verificação, constatou-se que os clubes os quais possuíam mais simpatizantes no Estado do Paraná eram o Athlético Paranaense, o Coritiba e o Paraná ambos clubes da capital paranaense e rivais. Considerou-se somente as pesquisas realizadas no Paraná devido ao fato de ser o local onde está sendo desenvolvido o presente projeto, facilitando a aproximação com os dados e sua coleta.

A pesquisa visa contribuir para uma inteligência esportiva mais ampla, atingindo novas formas de buscar, coletar e analisar dados, visando *insights* significativos através da opinião popular nas redes sociais. Clubes de futebol podem se aproveitar de conceitos de *Machine Learning*, Mineração de Textos e Recuperação da Informação para incrementar suas capacidades analíticas de marketing, imagem e valor.

Inicialmente, buscou-se o tema “Análise de sentimentos” acrescido dos termos “Futebol” e “Twitter” no acervo da Universidade Federal do Paraná (UFPR), com o objetivo de iniciar o processo de estudo na área e analisar as pesquisas já realizadas na área. A verificação do tema foi realizada durante o mês de maio de 2019. A pesquisa retornou 40 resultados disponíveis em formato PDF para leitura e consulta. Os 40 artigos retornados foram analisados separadamente, buscando encontrar relação entre este projeto de pesquisa e a produção científica já existente. Após a análise individual das publicações retornadas da pesquisa, constatou-se que nenhum artigo possuía íntima relação com os assuntos “análise de sentimentos”, “twitter” e “futebol”.

A partir dos resultados da pesquisa realizada no acervo da UFPR, foi realizada uma revisão bibliográfica na base de dados Scopus, no portal de periódicos da CAPES, com o intuito de verificar as pesquisas realizadas na área de análise de sentimentos no futebol utilizando a rede social Twiiter como base para a análise. Inicialmente, foi buscado o descritor “*Sentiment Analysis”,* o qual retornou um total de 10.156 documentos, evidenciando o interesse pela área de análise de sentimentos. Em seguida foi pesquisado o termo “*Sentiment Analysis”* juntamente com a palavra “*Twitter”,* retornando 2.404 documentos.

Com os documentos retornados através da pesquisa feita com os termos "*Sentiment Analysis"* e "Twiiter", foi executada uma análise bibliométrica para maior aprofundamento do tema.

Os 2.404 documentos recuperados são de um período entre 2009-2020, com um total de 5.398 autores e 1.041 fontes diferentes. Além disso, destacam-se os documentos publicados em conferências e artigos, conforme expõe o quadro 1.

QUADRO 1 - TIPOS DOS DOCUMENTOS RETORNADOS PELA PESQUISA BIBLIOMÉTRICA

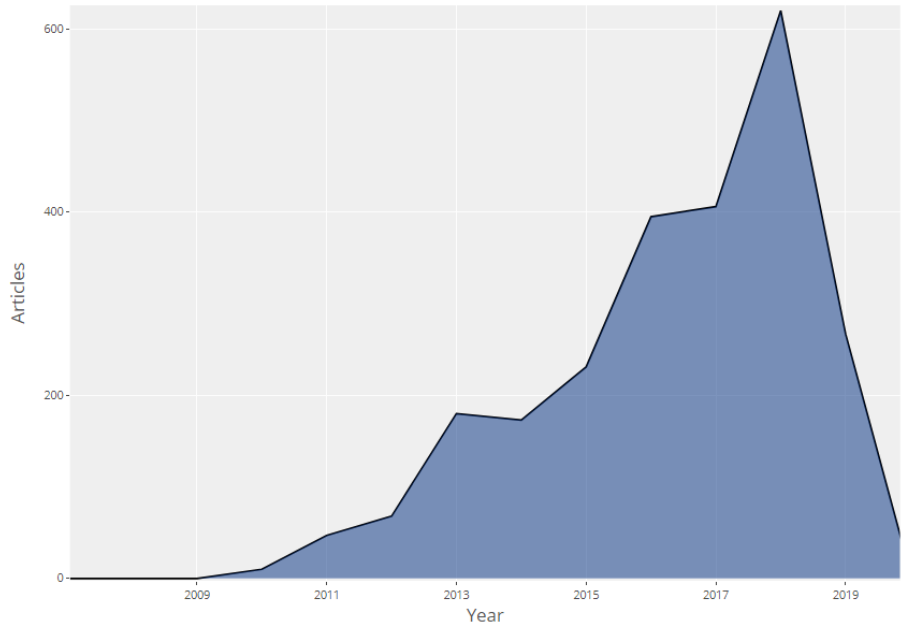
| **Tipo do documento** | **Quantidade** |
|---|---|
| Documentos publicados em conferências | 1.557 |
| Artigos | 671 |
| Revisões realizadas em conferências | 110 |
| Capítulo de livro | 44 |
| Revisões | 20 |
| Livro | 1 |
| Pesquisa rápida | 1 |

FONTE: O autor (2019)

Dessa forma, foi realizada uma análise da produção científica durante o período retornada pela pesquisa, a fim de observar a evolução do tema ao longo do tempo. A figura 1 exibe a produção científica durante o período.

A partir da figura 1, percebe-se o crescimento do tema principalmente após o ano de 2010, com seu ápice entre 2017 e 2018. Os artigos recuperados pela pesquisa datam a partir do ano de 2009 pelo fato do Twitter só ter sido criado em 2006, conforme afirma Smaal (2010): "O Twitter foi fundado em março de 2006 por Jack Dorsey, Evan Williams e Biz Stone como um projeto paralelo da Odeo". O crescimento com o passar dos anos pode ser explicado pela popularidade que as redes sociais ganharam com o tempo, tornando-se redes de comunicação global.

FIGURA 1 - PRODUÇÃO CIENTÍFICA ENTRE 2009 - 2019




FONTE: O autor (2019)

A partir das referências dos documentos, obteve-se as principais fontes citadas pelos pesquisadores, como parte de uma investigação a respeito das fontes mais populares sobre o tema. O quadro 2 apresenta as principais fontes citadas e a quantidade de vezes que os autores utilizaram a fonte.

QUADRO 2 - AS 10 FONTES MAIS CITADAS PELOS PESQUISADORES

| | |
|---|---|
| EXPERT SYSTEMS WITH APPLICATIONS | 378 |
| JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY | 294 |
| ICWSM | 286 |
| PLOS ONE | 251 |
| DECISION SUPPORT SYSTEMS | 244 |
| FOUNDATIONS AND TRENDS IN INFORMATION RETRIEVAL | 234 |
| EXPERT SYST APPL | 230 |
| COMPUTATIONAL LINGUISTICS | 197 |
| TWITTER SENTIMENT CLASSIFICATION USING DISTANT SUPERVISION | 165 |
| JOURNAL OF MACHINE LEARNING RESEARCH | 162 |

FONTE: O autor (2019)

O mesmo procedimento foi executado para saber os autores mais citados, bem como o nome do trabalho mais popular dos resultados obtidos através da consulta. O quadro 3 demonstra os documentos citados e o número de citações que este documento recebeu.

QUADRO 3 - DOCUMENTOS COM O MAIOR NÚMERO DE CITAÇÕES

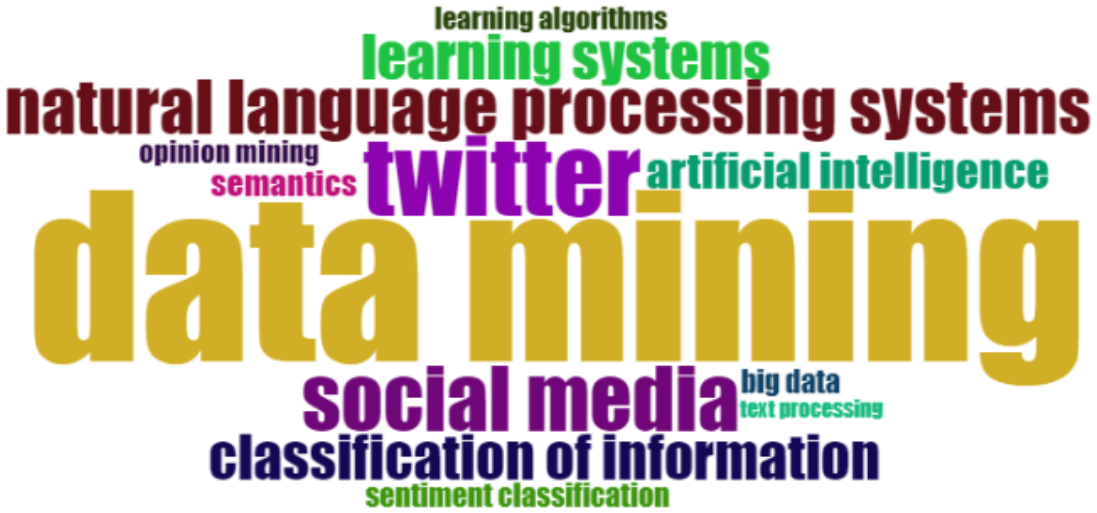
| | |
|---|---|
| PANG, B., LEE, L., OPINION MINING AND SENTIMENT ANALYSIS (2008) FOUNDATIONS AND TRENDS IN INFORMATION RETRIEVAL, 2 (1-2), PP. 1-135 | 126 |
| BOLLEN, J., MAO, H., ZENG, X., TWITTER MOOD PREDICTS THE STOCK MARKET (2011) JOURNAL OF COMPUTATIONAL SCIENCE, 2 (1), PP. 1-8 | 80 |
| LIU, B., SENTIMENT ANALYSIS AND OPINION MINING (2012) SYNTHESIS LECTURES ON HUMAN LANGUAGE TECHNOLOGIES, 5 (1), PP. 1-167 | 71 |
| TABOADA, M., BROOKE, J., TOFILOSKI, M., VOLL, K., STEDE, M., LEXICON-BASED METHODS FOR SENTIMENT ANALYSIS (2011) COMPUTATIONAL LINGUISTICS, 37 (2), PP. 267-307 | 55 |
| JANSEN, B.J., ZHANG, M., SOBEL, K., CHOWDURY, A., TWITTER POWER: TWEETS AS ELECTRONIC WORD OF MOUTH (2009) JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY, 60 (11), PP. 2169-2188 | 46 |
| THELWALL, M., BUCKLEY, K., PALTOGLOU, G., SENTIMENT IN TWITTER EVENTS (2011) JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY, 62 (2), PP. 406-418 | 38 |
| MEDHAT, W., HASSAN, A., KORASHY, H., SENTIMENT ANALYSIS ALGORITHMS AND APPLICATIONS: A SURVEY (2014) AIN SHAMS ENGINEERING JOURNAL, 5 (4), PP. 1093-1113 | 36 |
| THELWALL, M., BUCKLEY, K., PALTOGLOU, G., CAI, D., KAPPAS, A., SENTIMENT STRENGTH DETECTION IN SHORT INFORMAL TEXT (2010) JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY, 61 (12), PP. 2544-2558 | 34 |
| FELDMAN, R., TECHNIQUES AND APPLICATIONS FOR SENTIMENT ANALYSIS (2013) COMMUNICATIONS OF THE ACM, 56 (4), PP. 82-89 | 29 |
| THELWALL, M., BUCKLEY, K., PALTOGLOU, G., SENTIMENT STRENGTH DETECTION FOR THE SOCIAL WEB (2012) JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY, 63 (1), PP. 163-173 | 28 |

FONTE: O autor (2019)

Com o intuito de obter as principais palavras utilizadas pelos autores nos documentos recuperados e exibi-las de forma visual, criou-se uma nuvem de palavras com os termos mais populares. A figura 2 retrata a nuvem de palavras com os termos mais frequentes recuperados pela pesquisa.

Nota-se a partir da figura 2 a popularidade do termo "*data mining*", também conhecido como mineração de dados. A mineração de dados segundo Castro e Ferrari (2016, p. 7) é uma disciplina multidisciplinar que envolve áreas como estatística, aprendizagem de máquina, recuperação da informação e inteligência artificial. Para os autores, a mineração de dados corresponde à aplicação de algoritmos capazes de extrair conhecimentos a partir dos dados. A alta menção ao termo pode ser explicada pelo fato de a mineração de dados ser a base para as técnicas de mineração de texto, conforme afirmam Aranha e Passos (2016).

FIGURA 2 - NUVEM DE PALAVRAS COM OS TERMOS MAIS POPULARES




FONTE: O autor (2019)

Após a análise bibliométrica realizada com os termos descritores "*Sentiment Analysis"* e "Twitter", com o intuito de filtrar ainda mais os resultados de acordo com o tema deste projeto de pesquisa, foi acrescida a palavra em inglês para futebol ("*soccer"),* retornando oito documentos no total. Destes 8 documentos, somente 3 tinham relação com o tema, conforme apresenta o quadro 4.

QUADRO 4 - DOCUMENTOS RELACIONADOS COM O PROJETO DE PESQUISA

| Título do Documento | Autores | Ano de Publicação |
|---|---|---|
| Soccer events summarization by using sentiment analysis | Jai-Andaloussi, S., El Mourabit, I., Madrane, N., Chaouni, S.B., Sekkaki, A. | 2016 |
| World Cup 2014 in the Twitter World: A big data analysis of sentiments in U.S. sports fans' tweets | Yu, Y., Wang, X. | 2015 |
| Sentiment Identification in Football-Specific Tweets | Aloufi, S., Saddik, A.E. | 2013 |

FONTE: O Autor (2019)

A carência de estudos relacionando conceitos de análise esportiva, mineração de textos e redes sociais foi um motivador para esta pesquisa, a qual busca ser um

estudo para contribuir para o possível desenvolvimento de novos estudos relacionados com a área. Além disso, este trabalho em pauta pretende exibir de forma prática como tais conceitos podem impactar de maneira direta as estratégias de um clube, podendo ser útil para que os clubes tenham interesse em explorar as redes sociais como uma fonte de recuperação das informações através de técnicas de aprendizado de máquina, incrementado suas capacidades analíticas.

A presente pesquisa utiliza conceitos estudados durante o curso de Gestão da Informação, tais como recuperação da informação, mineração de dados e aprendizado de máquina para encontrar informações relevantes, capazes de gerar conhecimento e consequentemente, auxiliar na tomada de decisões. O tema deste projeto de pesquisa pode contribuir para evidenciar conceitos estudados ao longo do aprendizado, podendo servir como exemplo para novos estudos na área de recuperação da informação, análise de sentimentos e mineração de opiniões.

Vale ressaltar que o assunto “mineração de opiniões” já foi estudado no curso de Gestão da Informação por outros egressos, mas com enfoques diferentes desta pesquisa. Rodrigues (2017) estuda análise de sentimentos na rede social Facebook, proporcionando um modelo para analisar os sentimentos contidos na página do Senado Federal Brasileiro. Já Iglesias (2018) explora a mineração de opiniões no Serviço de Atendimento ao Consumidor (SAC) através da página do Facebook, com o intuito de medir a insatisfação do consumidor. Outro estudo relacionado com a mineração de opiniões foi desenvolvido por Foit (2017), o qual verifica a opinião dos usuários a respeito do assunto Bitcoin no Twitter. Tanto Rodrigues (2017) quanto Iglesias (2018) utilizam de algoritmos de aprendizado de máquina para realizar a classificação e análise dos sentimentos, enquanto Foit (2017) utiliza o *software* Microsoft Excel ® na classificação das opiniões coletadas.

Os projetos anteriormente citados evidenciam a proximidade do curso de Gestão da Informação com o tema em pauta, contribuindo para o desenvolvimento e estudo do assunto de mineração de opiniões.

## 1.4 DELIMITAÇÃO DA PESQUISA

A presente pesquisa visa explorar o tema recuperação da informação juntamente com análise de sentimentos nas redes sociais e, ao final construir exibir resultados obtidos. Este projeto de pesquisa delimitou-se em coletar e analisar os

sentimentos dos torcedores dos três clubes de futebol do Estado do Paraná com a maior quantidade de simpatizantes, Athlético Paranaense, Coritiba e Paraná, conforme apresenta Rudnick (2017).

A pesquisa em pauta limita-se somente em recuperar informações somente da rede social Twitter durante o período entre 08/09/2019 e 19/10/2019. A escolha da rede social ocorreu devido ao fato de ser um microblog caraterizado por postagens curtas, com no máximo 280 caracteres, estimulando o envio de mensagens objetivas e diretas. Outro fator determinante para a escolha da rede social está no fornecimento de *Application Programming Interface* (API), interfaces de programação de aplicações que permitem acessar um grande volume de informações a respeito de atualizações, status e dados dos usuários de forma facilitada. As APIs disponíveis no Twitter viabilizam o acesso à uma grande quantidade de informações, executando tal processo de forma automatizada e dinâmica.

## 1.5 ESTRUTURA DO DOCUMENTO

O documento está dividido em cinco partes. Na primeira seção é apresentada a introdução, problematização e questão de pesquisa, objetivos gerais e específicos, justificativa e a delimitação da pesquisa.

A segunda seção contempla a revisão de literatura relacionada à gestão da Informação, recuperação da informação, aprendizado de máquina, mineração de texto, análise de dados no esporte, gestão de clubes de futebol, relação entre os conceitos de mineração de texto e futebol e análise de sentimentos. Além disso, esta seção apresenta a rede social Twitter.

A seção três trata da metodologia proposta para a execução desta pesquisa, e criação da base de dados, além dos encaminhamentos metodológicos que foram utilizados no pré-processamento e na análise de sentimentos para alcançar o objetivo traçado

Na quarta seção estão os resultados obtidos com a execução das atividades desenvolvidas ao longo do projeto e as análises realizadas.

A quinta seção trata das considerações finais, trazendo as contribuições do estudo realizado e possíveis pesquisas futuras relacionadas.

# 2 REVISÃO DE LITERATURA

Neste capítulo, são apresentados conceitos relevantes para suportar a problemática pautada e os objetivos traçados, tais como definições sobre aprendizado de máquina, mineração de textos, processamento de linguagem natural, gestão de dados no esporte, recuperação da informação e uma elucidação a respeito da rede social utilizada nesta pesquisa, o Twitter.

## 2.1 GESTÃO DA INFORMAÇÃO

Segundo Davenport (1998, p. 173) o gerenciamento informacional ou gestão da informação “trata-se de um conjunto estruturado de atividades que incluem o modo como às empresas obtêm, distribuem e usam a informação e o conhecimento”. O autor prossegue afirmando que o gerenciamento da informação pode ser definido como um processo, o qual traz consigo métodos, ferramentas e técnicas orientadas para a informação.

A definição de informação por Davenport (1998, p.18) é apresentada através de uma abordagem definindo inicialmente os conceitos de dados, informação e conhecimento, conforme ilustra o quadro 5:

QUADRO 5 - DISTINÇÃO ENTRE DADO, INFORMAÇÃO E CONHECIMENTO

| Dados | Informação | Conhecimento |
|---|---|---|
| Simples observações sobre o estado do mundo<br>Facilmente estruturado<br>• Facilmente obtido por máquinas<br>• Freqüentemente quantificado<br>• Facilmente transfcrível | Dados dotados de relevância e propósito<br>• Requer unidade de análise<br>• Exige consenso em relação ao significado<br>• Exige necessariamente a mediação humana | Informação valiosa da mente humana<br>Inclui reflexão, síntese, contexto<br>• De difícil estruturação<br>• De difícil captura em máquinas<br>• Freqüentemente tácito<br>• De difícil transferência |

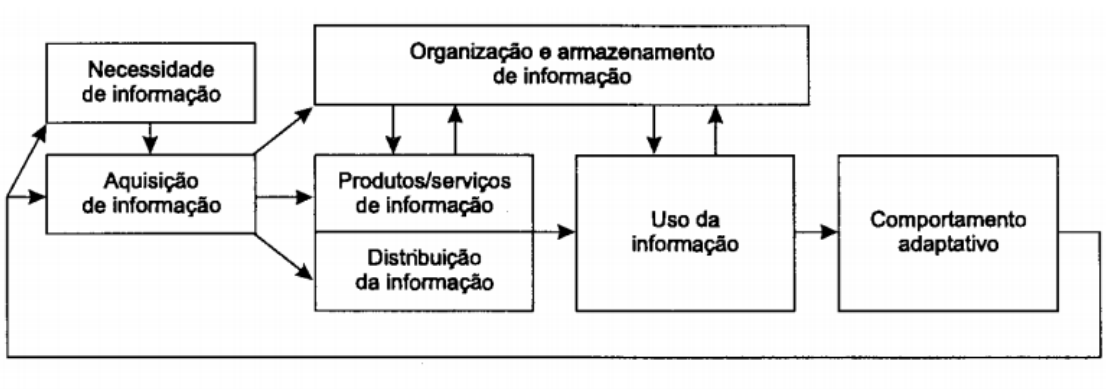
FONTE: Davenport (1998, p. 18)

Choo (2003, p. 27) destaca três usos distintos para a informação, as quais são chamadas de arenas de uso da informação e desempenham um papel estratégico. O primeiro uso da informação é com o intuito de dar sentido às mudanças do ambiente

externo, distinguindo as mudanças mais significativas, interpretando-as e criando respostas para elas. O segundo uso da informação está atrelado ao processo de gerar novos conhecimentos por meio do aprendizado, permitindo o desenvolvimento de novas capacidades. Por fim, o terceiro uso da informação relaciona-se ao processo de busca e avaliação de informações com o objetivo de tomar decisões importantes. O autor afirma que as três arenas de uso da informação são processos dinâmicos e interligados, os quais fornecem uma visão holística do uso da informação, além de constituírem significados, conhecimentos e ações.

Choo (2003, p. 403) analisa a gestão da informação a partir da criação de seis processos necessários para atender as necessidades, buscas e uso da informação: identificação das necessidades de informação; aquisição da informação; organização e armazenamento da informação; desenvolvimento de produtos e serviços de informação; distribuição da informação; e uso da informação. Tais conjuntos de processos formam um ciclo da informação, conforme apresenta a figura 3.

FIGURA 3 - CICLO DA GESTÃO DA INFORMAÇÃO




FONTE: Choo (2003, p. 404)

Para Choo (2003, p. 405) as necessidades de informação nascem de problemas, incertezas e ambiguidades encontradas em situações e experiências específicas, relacionadas à clareza dos objetivos. Existe uma preocupação não somente com o significado, mas também com as condições, padrões e regras de uso, que tornam a informação significativa para determinados indivíduos em determinadas situações.

A aquisição da informação segundo Choo (2003, p. 407) é uma função crítica e complexa dentro da gestão da informação, de modo que a seleção e o uso das fontes de informação têm de ser planejados e continuamente monitorados e avaliados,

como qualquer outro recurso vital para a organização. Ainda de acordo com o autor, as fontes para monitorar o ambiente devem ser suficientemente numerosas e variadas para refletir todo o espectro de fenômenos externos. O autor aborda uma maneira eficaz de administrar a variedade de informações, a qual se baseia em envolver o maior número possível de pessoas na coleta de informações, contribuindo para a filtragem das informações e oferecendo uma comunicação rica e satisfatória.

Choo (2003, p. 409) avalia a organização e o armazenamento da informação como um facilitador de recuperação e disseminação de informação. Segundo o autor, a maneira como a informação é armazenada em arquivos, bancos de dados computadorizados e outros sistemas de informação representa um componente importante e frequentemente consultado da memória da organização.

O processo de produtos e serviços informacionais segundo Choo (2003, p. 412) precisa abranger não apenas a área do problema, mas também as circunstâncias específicas que afetam a resolução de cada problema ou cada tipo de problema. Para o autor, produtos e serviços de informação são desenvolvidos como qualidades que agregam valor à informação que está sendo processada, com o objetivo de ajudar o usuário a tomar melhores decisões, a perceber melhor as situações e empreender ações mais eficazes.

Choo (2003, p. 414) estabelece a distribuição da informação como o processo pelo qual as informações se disseminam pela organização, de maneira que "a informação correta atinja a pessoa certa no momento, lugar e formato adequados". Conforme o autor explicita, o objetivo da distribuição da informação é promover e facilitar a partilha de informações, sendo algo fundamental para a criação de significado, a construção de conhecimento e a tomada de decisões.

Ainda segundo o autor, a informação é buscada e usada em todo o processo de tomada de decisões, devendo ser compartilhada facilmente, mas sem perda da riqueza cognitiva: “O uso da informação resulta da criação de significado, de conhecimento e de decisões. Em cada caso, o uso da informação é um processo social de pesquisa fluido, recíproco e repetitivo. ” (CHOO, 2003, p. 415).

Em outra abordagem, Assis (2008, p. 6) afirma que um bom gerenciamento da informação deve ser baseado em políticas que prevejam critérios de seleção e guarda, normas para organização e categorização, padronização, incentivo à disseminação e ao uso de informações, com amplo e democrático acesso.

Segundo Assis (2008, p. 141) um dos objetivos da gestão da informação é garantir que a informação seja administrada como um recurso valioso e indispensável, atestando que as informações estejam alinhadas com os objetivos e metas do negócio. O autor também elenca condições a serem observadas na implantação de uma gestão da informação, sendo elas:

- a gestão da informação deve estar alinhada com a missão e os objetivos estratégicos;
- desenvolver um plano de gestão da informação voltado, preferencialmente, para a perspectiva do negócio;
- preocupar-se sempre com a máxima: a informação para as pessoas certas, no local correto, no tempo certo, no formato adequado e, se possível, com custo zero;
- ter a visão de que a informação deve ser utilizada no seu potencial máximo;
- priorizar a qualidade, a disponibilidade, o uso e o valor da informação;
- o gestor da informação deve estar ligado diretamente à alta administração;
- mapear regularmente as necessidades de informação;
- considerar a qualidade das fontes de informação e sua disponibilidade;
- analisar o custo x benefício das fontes de informação adquiridas;
- contextualizar e compartilhar a informação de interesse.

A partir dos diversos estudos a respeito da Gestão da Informação, neste projeto de pesquisa será considerado o enfoque na recuperação da informação, tendo em vista a importância do tema para a análise de sentimentos nas redes sociais.

## 2.2 RECUPERAÇÃO DA INFORMAÇÃO

De acordo com Baeza-Yates e Ribeiro-Neto (1999, p. 1), a recuperação da informação lida com a representação, o armazenamento, a organização e o acesso às informações. Segundo os autores, a recuperação da informação trata principalmente com textos em linguagens naturais, ou seja, sem uma estruturação bem definida e com ambiguidade.

Baeza-Yates e Ribeiro-Neto (1999, p. 2) exibem a diferença da recuperação de dados para a recuperação da informação. Para os autores, a recuperação de dados

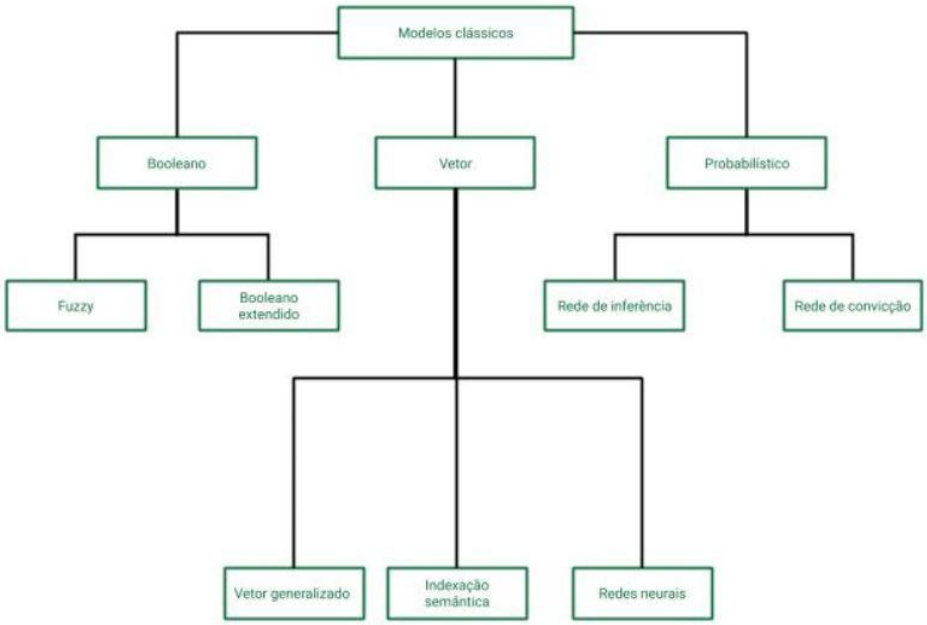
foca em soluções para usuários de banco de dados, sem a preocupação com o conteúdo dos dados, enquanto a recuperação da informação preocupa-se em interpretar o conteúdo e a relevância das mensagens, buscando resolver a solicitação dos usuários. O objetivo principal da recuperação da informação é recuperar todos os documentos pertinentes para os usuários, envolvendo a semântica e a sintaxe de um documento para obter a informação (BAEZA-YATES; RIBEIRO-NETO, 1999, p. 2).

Assis (2008, p. 143) menciona que “a recuperação da informação, em decorrência do crescimento e do grande volume de publicações existentes no mundo, tornou-se necessária para suprir as necessidades dos usuários. ”

Conforme abordam Baeza-Yates e Ribeiro-Neto (1999, p. 19), sistemas de recuperação da informação tradicionais apresentam dificuldades na seleção de quais documentos são relevantes e quais não são. A realização da tarefa de seleção dos documentos geralmente é realizada por algoritmos, os quais selecionam os documentos através de suas premissas previamente definidas (BAEZA-YATES; RIBEIRO-NETO, 1999, p. 19).

Três modelos clássicos de recuperação da informação são apontados por Baeza-Yates e Ribeiro-Neto (1999, p. 20), conhecidos como recuperação booleana, recuperação por vetores e recuperação probabilística. No modelo booleano os documentos são representados por um conjunto de termos indexados, sendo um modelo conhecido como teórico. O modelo de recuperação por vetores utiliza vetores em um espaço tridimensional para representar os documentos, sendo conhecido como um modelo algébrico. Já o modelo probabilístico utiliza a teoria da probabilidade para estruturar e caracterizar os documentos. Conforme os estudos avançaram na área de recuperação da informação, novos modelos foram propostos a partir dos modelos clássicos (BAEZA-YATES; RIBEIRO-NETO, 1999, p. 20). A figura 4 elenca os modelos clássicos e suas principais derivações.

FIGURA 4 - MODELOS CLÁSSICOS DE RECUPERAÇÃO DA INFORMAÇÃO




FONTE: Adaptado de BAEZA-YATES E RIBEIRO-NETO (1999, p. 21)

Baeza-Yates e Ribeiro-Neto (1999, p. 73) comentam que a implementação de um sistema de recuperação depende do objetivo do sistema, algo que deve ser analisado previamente, verificando suas especificidades e funcionalidades.

Cunningham, Littin e Witten (1997, p. 2) afirmam que a recuperação da informação possui múltiplas características da área de aprendizado de máquina. Segundo os autores, algoritmos de aprendizado de máquina utilizam exemplos, atributos e valores fornecidos em abundância por sistemas de recuperação da informação. Conforme argumentam Cunningham, Littin e Witten (1997, p. 2) “a grande quantidade de documentos nos sistemas de recuperação da informação permite a utilização de uma infinidade de recursos de linguagem natural, fornecendo uma riqueza de peculiaridades”.

O processo de recuperação da informação pode ser dividido em quatro fases distintas: indexação, formulação de consulta, comparação e feedback, as quais fornecem oportunidades para a utilização de técnicas de aprendizado de máquina (CUNNINGHAM; LITTIN; WITTEN, 1997, p. 7). De acordo com os autores, uma das maneiras de utilizar o aprendizado de máquina nos processos de recuperação da informação consiste na representação de textos baseados em modelos vetoriais, também conhecido como “*bag of words”*, caracterizado por representar um documento

através de um conjunto de termos singulares. Cunningham, Littin e Witten (1997, p. 8) elucidam que na técnica de *bag of words,* os documentos contêm um determinado conjunto de termos, os quais correspondem a um ponto estabelecido no espaço n-dimensional, com n referindo-se ao número de termos. Na interpretação dos autores, estes termos são geralmente palavras, e um vetor de termo é um vetor booleano que representa o conjunto de palavras que aparecem no documento ou um vetor numérico cujos valores são derivados do número de ocorrências de cada termo em um documento.

Cunningham, Littin e Witten (1997, p. 8) explanam a respeito da utilização do mecanismo *bag of words*, o qual baseia-se no entendimento de que quanto mais frequente um termo é mencionado, maior a probabilidade deste termo estar conectado ao assunto do documento. O cálculo vetorial do documento ocasiona a perda de todas as outras informações implícitas no texto, como o significado e a ordem das palavras, no entanto, a simplicidade deste modelo torna-o um dos mais populares no campo da recuperação da informação (CUNNINGHAM; LITTIN; WITTEN, 1997, p. 8).

Outra aplicação das técnicas de aprendizado de máquina dentro da recuperação da informação reside na categorização de textos. A categorização de textos é “o processo de classificar os documentos em uma categoria específica” (CUNNINGHAM; LITTIN; WITTEN, 1997, p. 18). O aprendizado de máquina permite a automação da categorização dos documentos, o que aumenta a consistência da classificação, além da diminuição do tempo e dos gastos. Dentre os diversos algoritmos de aprendizado de máquina utilizados em sistemas de recuperação da informação, destacam-se: máquinas de vetores de suporte (SVM), redes neurais, classificadores bayesianos e árvores de decisão (CUNNINGHAM; LITTIN; WITTEN, 1997, p. 20).

Neste projeto será utilizada a recuperação da informação para extrair informações dos dados recuperados no Twitter e categoriza-los de acordo com o sentimento contido nele. A partir dos conceitos evidenciados pela área de recuperação da informação, faz-se necessária a apresentação da área de aprendizado de máquina, uma vez que diversas técnicas deste âmbito são utilizadas na recuperação da informação, conforme apresentado anteriormente.

## 2.3 APRENDIZADO DE MÁQUINA

*Machine Learning* ou aprendizado de máquina (AM) é o processo de converter a experiência em conhecimento a partir de um determinado algoritmo de entrada. A entrada de um algoritmo de aprendizado é o treinamento de dados, representando a experiência, enquanto a saída é feita geralmente através de um programa de computador capaz de executar uma tarefa, representando o conhecimento. O aprendizado de máquina é uma área interdisciplinar, utilizando conceitos de estatística, ciência da computação e inteligência artificial para detectar padrões significativos em bases de dados ou partir da interação com o ambiente (SHALEV-SHWARTZ; BEN-DAVID, 2014, p. 24).

Shalev-Shwartz e Ben-David (2014, p. 21) ressaltam que o AM é utilizado quando as atividades são muito complexas para programas usuais de computadores ou quando é necessária a adaptação dos programas, que reagem às mudanças no ambiente com o qual eles interagem.

Vanderplas (2017, p. 332) define a área de aprendizado de máquina como responsável pela construção de modelos matemáticos capazes de auxiliar na compreensão dos dados. O autor afirma que o aprendizado ocorre quando os modelos matemáticos recebem parâmetros ajustáveis que podem ser adaptados a partir dos dados, permitindo ao programa aprender a partir dos dados.

De acordo com Simon (1983 apud SATHYA; ABRAHAM, 2013) o aprendizado de máquina pode se referir tanto à aquisição quanto ao aprimoramento do conhecimento. A aprendizagem de máquina denota mudanças adaptativas, permitindo um sistema realizar a mesma tarefa várias vezes, mas aumentando sua efetividade e eficiência a medida do tempo.

Murphy (2012, p. 2) aponta dois tipos principais de aprendizado de máquina, o aprendizado supervisionado e o aprendizado não supervisionado. Conforme cita o autor, no aprendizado supervisionado ou preditivo o objetivo é aprender um mapeamento de determinada entrada x para saídas y, dado um conjunto de pares de entrada-saída (x, y), sendo comumente empregado em atividades como classificação de documentos, filtro de spams em e-mails, categorização de flores, reconhecimento de imagens e escritas e na detecção facial. Já na abordagem não supervisionada ou descritiva o objetivo é descobrir estruturas interessantes nos dados, porém, sem qualquer informação a respeito da saída desejada.

Kelleher, Namee e D'arcy (2015, p. 7) destacam que os algoritmos de aprendizado de máquina funcionam verificando um conjunto de possíveis modelos de previsão para o modelo que melhor captura a relação entre os recursos descritivos e o recurso de destino em um conjunto de dados. Segundo os autores, um critério para conduzir essa busca é procurar modelos que são consistentes com os dados.

Castro e Ferrari (2016, p. 50) afirmam que sistemas de aprendizagem são aqueles capazes de se adaptar ou mudar seu comportamento com base em exemplos, de forma que manipule informações. Para os autores, duas virtudes importantes da aprendizagem são a possibilidade de resolver tarefas de processamento de informação e a capacidade de operar em ambientes dinâmicos. De acordo com os autores, "a maioria dos processos de aprendizagem é gradativa, ou seja, a aprendizagem não ocorre instantaneamente, requerendo um processo interativo e/ou iterativo de adaptação e interação com o ambiente. "

Conforme Castro e Ferrari (2016, p. 51) elencam, o aprendizado de máquina tem como foco extrair informação a partir de dados de maneira automática, estando intimamente relacionada à mineração de dados, à estatística, à inteligência artificial e à teoria da computação, além de outras áreas como computação natural, sistemas complexos adaptativos e computação flexível.

Dentre as diversas aplicações utilizadoras de princípios de aprendizado de máquina, este presente trabalho irá enfatizar a mineração de textos, posto que a área está intimamente conectada aos conceitos de recuperação da informação e redes sociais.

## 2.4 MINERAÇÃO DE TEXTOS

Segundo Carvalho Filho (2014, p.16 apud Hearst, 1999) "dados textuais englobam uma vasta e rica fonte de informação, mesmo em um formato que seja difícil de extrair de maneira automatizada". Assim, são necessárias aplicações de métodos e algoritmos para dar estruturação aos dados textuais, objetivando facilitar a extração de conhecimento dos respectivos dados, processo este conhecido como Mineração de Textos.

De acordo com Tan, Mui e Terrace (1999, p. 1) "a mineração de texto ou mineração de dados textuais refere-se ao processo de extrair padrões ou conhecimentos interessantes a partir de documentos textuais não estruturados".

Aranha e Passos (2006) afirmam que a mineração de textos é um conjunto de métodos usados para navegar, organizar, achar e descobrir informação em bases textuais, podendo ser enxergada como uma extensão da área de Mineração de Dados focada na análise de textos. Os autores ainda discorrem sobre o crescimento do armazenamento de dados não estruturados devido ao avanço da mídia digital, fato que propiciou o desenvolvimento das técnicas de mineração de textos.

Zanini e Dhawan (2015, p. 38) definem a mineração de textos como um conjunto de técnicas estatísticas e de ciência da computação para analisar dados em formatos de textos. De acordo com os autores, textos sempre foram fontes informativas que com o avanço das tecnologias passaram a ser estudados com o objetivo de extrair informações através de sistemas automáticos.

A mineração de texto segundo Zanini e Dhawan (2015, p. 38) é uma combinação de diversos campos relacionados, tais como mineração de dados, inteligência artificial, estatística, gestão de dados, bibliotecas cientificas e linguística. O objetivo básico da área de mineração de texto para os autores é processar a informação não estruturada contida nos dados de texto para tornar o texto acessível a vários algoritmos estatísticos de mineração de dados, permitindo investigar relações e padrões que de outra forma seriam extremamente difíceis de descobrir.

A partir da mineração de textos as informações contidas nos documentos podem ser categorizadas e agrupadas com o objetivo de produzir resultados como distribuição de frequência de palavras, padrões de reconhecimento e análise preditiva, podendo ser uma fonte estratégica de informações baseadas em evidências (ZANINI; DHAWAN, 2015, p. 38).

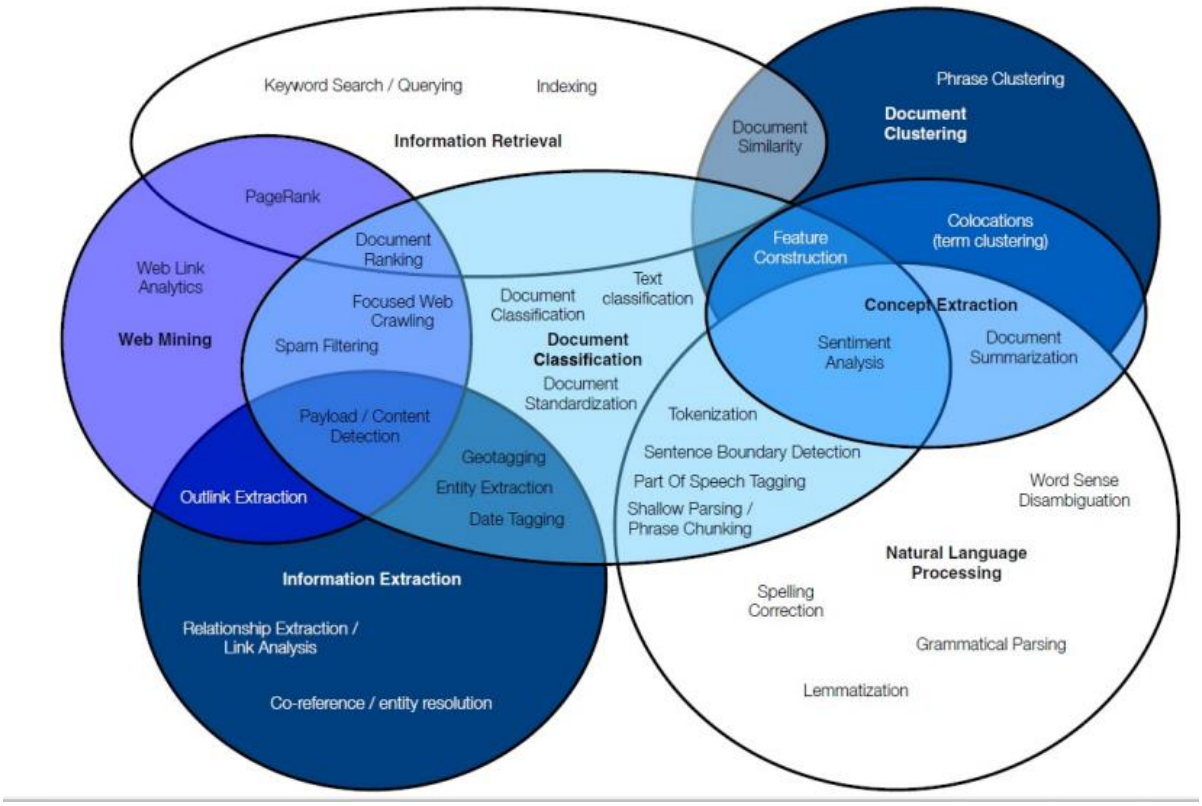
Talib et al. (2016, p. 414) discorrem a respeito das áreas de aplicação das técnicas de mineração de textos, utilizadas para mecanismos de pesquisa, sistema de gerenciamento de relacionamento com clientes, e-mails de filtro, análise de sugestões de produtos, detecção de fraudes e análise de redes sociais.

Em redes sociais são empregados softwares de mineração de texto, os quais assessoram na identificação e na análise do número de publicações, comentários e interações nas redes sociais. Este tipo de análise informa o comportamento dos usuários em diferentes postagens e auxilia no monitoramento das publicações (TALIB et al., 2016, p. 417)

Talib et al. (2016, p. 415) elucidam sobre as diferentes técnicas de mineração de texto que são aplicadas para analisar os padrões de texto e o processo de

mineração, além de suas inter-relações com outras áreas. A figura 5 corrobora com a afirmação dos autores.

FIGURA 5 - RELAÇÃO ENTRE AS TÉCNICAS UTILIZADAS NA MINERAÇÃO DE TEXTOS




FONTE: Talib et al. (2016, p. 415)

De forma semelhante, Zanine e Dhawan (2015) consideram uma combinação de tarefas baseadas em quatro diferentes estágios na aplicação das técnicas de mineração de textos, sendo elas:

- recuperação da informação;
- processamento de linguagem natural (PLN);
- extração da informação;
- mineração de dados.

Conforme apresenta Zanini e Dhawan (2015, p. 38), o primeiro estágio da mineração de textos é identificar documentos relevantes através de sistemas de recuperação de informação, visando corresponder à consulta de um usuário.

A segunda etapa para Zanini e Dhawan (2015, p.39) é realizada após a coleta dos documentos, utilizando-se de técnicas de PLN nos caracteres apresentados nos

documentos, com a finalidade de processá-los para serem analisados por computadores. Nesta fase, os documentos ainda estão em uma linguagem não-estruturada.

Após a etapa de PLN ocorre a extração da informação, que identifica os termos específicos gerados pelo estágio anterior e estrutura-o para a etapa de mineração de dados. A extração da informação permite vincular relacionamentos entre determinados eventos ou entidades (ZANINI; DHAWAN, 2015, p. 39).

Com os dados em formato estruturado ocorre a última etapa do processo de mineração de textos, que Zanini e Dhawan (2015, p. 39) conceituam como a captação de informações úteis dos dados textuais para construir novos conhecimentos. A mineração de textos é realizada com auxílio de técnicas e procedimentos estatísticos aplicados aos dados.

Barion e Lago (2008, p. 139) especificam a diferença entre a mineração de texto e mecanismos de busca. Segundo os autores, a mineração de texto auxilia o usuário na descoberta de informações desconhecidas enquanto na busca o usuário já tem conhecimento do que deseja encontrar. Ainda de acordo com Barrion e Lago (2008, p. 139) “como muitas informações estão armazenadas em formas de texto (mais de 80%), as técnicas de mineração de textos são muito importantes para a recuperação do conhecimento implícito”.

A partir das definições exibidas pelos autores, este projeto de pesquisa utilizará os conceitos de mineração de textos para extrair informações valiosas através dos dados coletados na rede social Twitter, conforme apresentado na seção de objetivos.

## 2.5 TWITTER

De acordo com Manrai et al. (2013, p. 237) as redes sociais eram muitas vezes ilegais e sem regulamentação, utilizadas por indivíduos com conhecimento tecnológico a fim de contornar, interromper ou copiar códigos, programas e serviços usados por empresas e indivíduos. Não foi amplamente adotado pelas empresas e não foi dado muito valor à utilidade das tecnologias. Conforme explicam os autores, a introdução da “World Wide Web (www) ” permitiu a proliferação de mídias sociais para um público muito mais amplo, evoluindo a forma como a comunicação e o compartilhamento de informações é feito.

Para Torres (2009, p. 113), as redes sociais são sites na Internet que permitem a criação e o compartilhamento de informações e conteúdo, sendo um espaço de produção e consumo de conteúdos aberto à colaboração e interação de modo geral. O autor aponta que o fato das redes sociais de serem colaborativas carregam diversas ferramentas de relacionamento, permitindo que as pessoas troquem mensagens, criem comunidades e se conheçam.

Torres (2009, p. 11) ainda apresenta que as redes sociais aglutinaram uma quantidade enorme de pessoas, as quais passaram a produzir informações e consumir conteúdos, tornando-as formadoras de opiniões.

Dentre as variadas redes sociais existentes na Internet está o Twitter, o qual foi fundado em 2006. O Twitter permite aos usuários enviar atualizações pessoais sobre o que estão fazendo, sendo utilizado para atualizações frequentes de informações (TORRES, 2009, p. 150).

Russel (2013, p. 7) afirma que o Twitter pode ser descrito como um serviço de microblog que permite a comunicação entre as pessoas através de mensagens curtas de 280 caracteres[1]. Na visão do autor, estas mensagens correspondem a pensamentos ou ideias, de forma que a rede social se assemelha a um serviço global de mensagens de textos em alta velocidade e gratuito.

O Twitter possui mais de 500 milhões de usuários registrados, com mais de 100 milhões interagindo ativamente de forma mensal, possibilitando atividades de marketing e publicidade, conforme aponta Russel (2013, p. 7). O autor cita a curiosidade humana e a necessidade por compartilhar ideias um dos fatores para o sucesso do Twitter:

> Quer seja uma paixão por fofocas de celebridades, um desejo de acompanhar um time de futebol favorito, um grande interesse em determinado assunto político ou um desejo de se conectar com alguém novo, o Twitter oferece oportunidades ilimitadas para satisfazer a curiosidade dos usuários (RUSSEL, 2013, p. 7).

Russel (2013, p. 5) expõe que o Twitter é uma rica fonte de dados para a mineração em redes sociais. Segundo o autor, a abertura para o consumo público, o fornecimento de interfaces de programação de aplicações limpas e bem documentadas, vastas ferramentas de desenvolvimento e amplo apelo aos usuários de todas as esferas da sociedade tornam esta rede social excelente para a aplicação

[1] https://g1.globo.com/tecnologia/noticia/twitter-aumenta-limite-para-280-caracteres.ghtml

de mineração de dados. “Os dados do Twitter são interessantes porque os tweets acontecem na velocidade do pensamento e estão disponíveis para consumo, pois acontecem quase em tempo real” (RUSSEL, 2013, p. 5).

## 2.6 ANÁLISE DE SENTIMENTOS

De acordo com Pang e Lee (2008) saber o que as pessoas pensam sempre foi uma informação importante durante nosso processo de tomada de decisões. Na visão dos autores, o advento da internet possibilitou descobrir as opiniões de um vasto conjunto de pessoas, de modo que os usuários anseiam e confiam nas opiniões dadas por outras pessoas, sendo uma das diversas razões para uma grande onda de interesse em sistemas capazes de lidar com as opiniões dos usuários.

Para Patel, Prabhu e Bhowmick (2015, p. 24), a análise de sentimentos rastreia, examina e avalia o humor do público usando técnicas de processamento de linguagem natural, sendo usada para sistemas de inteligência de negócios, a fim de analisar as opiniões do público em relação à sua marca e, consequentemente, implementar estratégias de mercado. Pang e Lee (2008, p. 7) elencam diversas aplicações da análise de sentimentos, tais como: aplicações relacionadas com resenhas em *websites,* aplicações com tecnologia de subcomponentes, inteligência empresarial e governamental, opiniões de eleitores em votações e tantas outras. Lin e Kolcz (2012, p. 800) afirmam que a análise de sentimentos foi amplamente disseminada no âmbito comercial, principalmente quando aplicado às mídias sociais, gerenciamento de marca e relacionamento com o cliente, bem como insights sobre o comportamento do consumidor no mercado.

Muitos fatores contribuíram para o desenvolvimento da área de análise de sentimentos, conforme explica Pange e Lee (2008, p. 4):

- surgimento de métodos de aprendizado de máquina no processamento de linguagem natural e recuperação de informações;
- disponibilidade de conjuntos de dados para que os algoritmos de aprendizado de máquina sejam treinados;
- desenvolvimento de sites de agregação de revisão e;
- realização dos desafios intelectuais e aplicações comerciais e de inteligência que a área oferece.

Pela primeira vez na história da humanidade, há uma quantidade massiva de dados nas mídias sociais na web, favorecendo o crescimento da análise de sentimentos, impactando não somente a área de PLN, mas também ciências administrativas, ciências políticas, ciências econômicas e ciências sociais, pois são diretamente afetadas pelas opiniões das pessoas (LIU, 2012).

De acordo com Liu (2012), os problemas de pesquisa na análise de sentimentos são baseados conforme o nível de granularidade, sendo divididos em três níveis. O primeiro nível é denominado nível de documento e visa classificar se um documento expressa um sentimento positivo ou negativo. O segundo nível, conhecido como nível de sentença, busca denominar o sentimento de cada sentença dentro do documento, classificando como positivo, negativo ou neutro. Este nível está fortemente relacionado com a subjetividade, a qual distingue sentenças que expressam informações objetivas de sentenças que expressam visões e opiniões subjetivas. O terceiro nível, denominado nível de entidade e aspecto, não se propõe a verificar as construções linguísticas (documentos, parágrafos, sentenças ou frases) mas as opiniões diretamente, com a ideia de que a opinião consiste em um sentimento direcionado a um alvo. Liu (2012) desenvolve o conceito de que uma opinião sem a identificação de seu alvo possui uso limitado, de maneira que perceber a importância do alvo auxilia no melhor entendimento do problema de análise de sentimentos. O principal objetivo neste nível de análise é descobrir sentimentos sobre as entidades e seus aspectos.

Liu (2012) aponta que em termos de mídias sociais, os pesquisadores trabalham principalmente com análise de sentimentos relacionados com análises de produtos ou serviços e com o Twitter, pois os tweets são curtos e geralmente, direcionados.

Liu, Li e Guo (2012, p. 1678) abordam que o Twitter oferece oportunidades para a pesquisa em mineração de dados e processamentos de linguagem natural devido ao seu alto número de postagens. Contudo, os autores apresentam as dificuldades de realizar a análise de sentimentos no Twitter, elucidando fatores como a ambiguidade das postagens, gírias, erros ortográficos, siglas e uma grande quantidade de dados sem rótulos ou com ruídos.

Conforme explana Liu, Li e Guo (2012, p. 1678), a análise de sentimentos no Twitter, também conhecida como *Twitter Sentiment Analysis (*TSA) utiliza métodos analisadores de sentimentos baseados principalmente em métodos semi-

supervisionados e supervisionados. Métodos de aprendizado semi-supervisionados na TSA são caracterizados por classificar os dados a partir de dados com rótulos ruidosos, como *emoticons* e hashtags, utilizando tais *emoticons* e *hashtags* para a classificação da polaridade do tweet.

Ainda para Liu, Li e Guo (2012, p. 1678), métodos de aprendizado supervisionado buscam classificar os sentimentos através de dados manualmente rotulados, utilizando algoritmos já evidenciados nesta pesquisa, tais como o Naive Bayes e o SVM. Os autores afirmam que métodos de aprendizado supervisionado são intensos e consomem muito tempo, uma vez que os dados precisam ser rotulados manualmente.

Para a execução desta pesquisa, a TSA será realizada utilizando métodos de aprendizado supervisionado, uma vez que segundo Liu (2012): “em problemas de classificação texto, qualquer método de aprendizado supervisionado existente pode ser aplicado”.

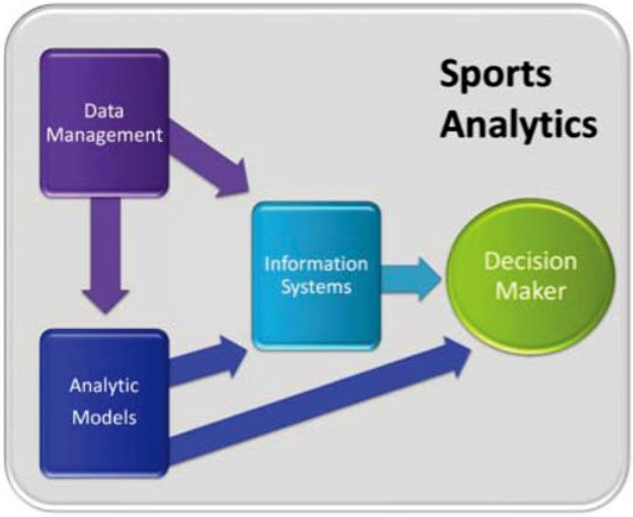
## 2.7 ANÁLISE DE DADOS NO ESPORTE

A análise esportiva segundo Alamar e Mehrotra (2011, p. 33) pode ser definida como:

> O gerenciamento de dados estruturados, a aplicação de modelos analíticos preditivos que utilizam estes dados e o uso de sistemas de informações para informar aos tomadores de decisão, permitindo auxiliar suas organizações a obter vantagem competitiva. (ALAMAR e MEHROTRA, 2011, p. 33).

A figura 6 auxilia na demonstração de uma estrutura de análise esportiva, conforme apontam os autores.

FIGURA 6 - ESTRUTURA DO MODELO DE ANÁLISE ESPORTIVA




Fonte: Alamar e Mehrotra (2011).

Conforme exemplifica a figura 6, a gestão dos dados é o primeiro elemento necessário para realizar uma análise esportiva, sendo tal gestão atrelada aos processos de coleta, verificação dos dados e o armazenamento de maneira eficiente. Dados de qualidade são essenciais para uma boa análise esportiva, já que os dados serão analisados através de modelos analíticos, além de serem utilizados nos sistemas de informações, os quais apoiam o processo decisório. Existe uma dificuldade em muitas organizações de obter os dados, pois muitas vezes eles não estruturados, dificultando o acesso. A falta de dados ou a inconsistência prejudicam a organização, reduzindo o valor da análise e seu impacto.

Após a gestão dos dados, são aplicadas ferramentas estatísticas, a fim de gerar *insights* fundamentais para a tomada de decisão e alimentar os sistemas de informações. A análise preditiva pode auxiliar na projeção da carreira de um jogador, na identificação de pontos fortes e fracos dos oponentes ou avaliar se um determinado jogador pode preencher uma necessidade da equipe por um preço justo.

A próxima etapa da estrutura consiste nos sistemas de informação, os quais estão sendo cada vez mais utilizados no mundo esportivo para a visualização de dados e a análise interativa com as informações. Os sistemas de informação fornecem uma plataforma de suporte à decisão, melhorando o modo como um tomador de decisão chega a uma ação.

Por fim, a última etapa da estrutura proposta por Alamar e Mehrotra (2011) é o tomador de decisão, o cliente final de todos os componentes da estrutura da análise esportiva.

A estrutura de análise esportiva engloba vários aspectos relacionados à transformação de dados brutos em informações que são valorizadas pelos tomadores de decisão no mundo dos esportes, conforme afirma Silva (2016):

> A estatística nos esportes é uma área em crescimento na estatística, a qual fornece uma metodologia especializada com o intuito de coletar e analisar dados esportivos para tomar decisões corretamente planejadas e implementar novas estratégias. (SILVA, 2016, p. 1). [2]

É importante abordar que as instituições esportivas modernas possuem muitos tomadores de decisão, como o gerente geral, treinadores, olheiros e outros executivos de pessoal. Tomadores de decisão em diferentes áreas funcionais podem utilizar dados e modelos diferentes para lidar com diferentes tipos de perguntas a serem solucionadas.

## 2.8 GESTÃO DE CLUBES DE FUTEBOL

Mattar e Mattar (2013) discorrem sobre as instituições esportivas e como as mesmas estão inseridas em um ambiente, pontuando à seguinte afirmação:

> As instituições esportivas estão inseridas em um ambiente de negócios que deve ser profundamente conhecido e analisado por seus gestores, permitindo um maior e melhor alinhamento entre os fatores e variáveis ambientais e seus planos, estratégias e ações, o que representa, em última análise, uma gestão mais assertiva (MATTAR e MATTAR, 2013, p. 7).

Os elementos que devem ser considerados por uma instituição esportiva de acordo com Mattar e Mattar (2013, p. 87) para uma definição adequada de sua estrutura organizacional são:

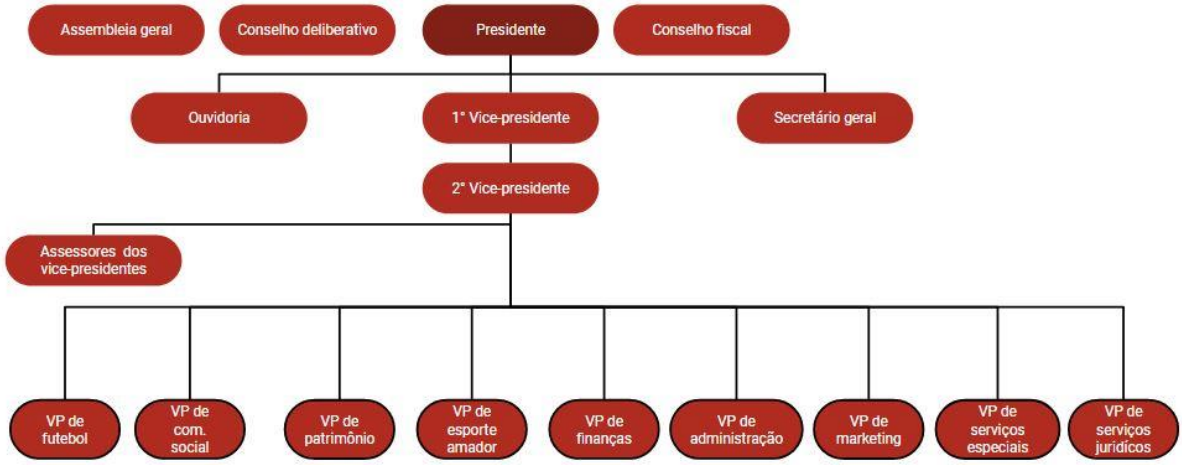
- as tarefas, os setores e as áreas necessárias para a operação;
- o grau de terceirização das atividades;
- o porte da organização.

[2] "Statistics in Sports is a growing field in Statistics that provides specialized methodology for collecting and analyzing sports data in order to make decisions for successful planning and implementation of new strategies.'

Para Mattar e Mattar (2013, p. 89), os clubes adotam um modelo tradicional de estrutura organizacional, com elevado nível de formalização, unidade de comando, comunicação vertical e alta departamentalização.

Figueiredo (2011, p. 191) apresenta um organograma de um clube de futebol, demostrando a estrutura existente na instituição, conforme apresentado na figura 7.

FIGURA 7 - EXEMPLO DE UMA ESTRUTURA ORGANIZACIONAL DE UM CLUBE DE FUTEBOL




Fonte: Adaptado de Figueiredo (2011, p. 192).

Conforme debatem Mattar e Mattar (2013, p. 92), a existência de uma estrutura organizacional projetada de maneira coordenada e integrada evita diversos problemas. Os autores elucidam os principais aspectos a serem considerados no delineamento de uma estrutura organizacional, sendo eles: critérios de departamentalização; amplitude de comando; número de níveis organizacionais; nível de descentralização; cadeia de comando; atribuição de responsabilidades; sistemas de comunicação; e sistema de detecção de problemas.

Dentre os setores apresentados pela figura 7, serão apresentados de modo aprofundado os setores mais próximos com o objetivo deste projeto de pesquisa, sendo eles: Planejamento estratégico, Gestão de marketing e Gestão financeira.

De acordo com Mattar e Mattar (2013, p. 95) o planejamento estratégico estabelece as diretrizes gerais e as estratégias que servirão como direcionamento para os planos, decisões e estratégias.

Mattar e Mattar (2013, p. 104) afirmam que as estratégias desenvolvidas pelos clubes buscam potencializar vantagens competitivas sustentáveis perante seus concorrentes. Uma das formas para atingir tal feito, segundo os autores, é reduzir os custos, pois custos mais baixos geram vantagens competitivas internas que possibilitam ao clube elevar a rentabilidade e melhorar a competitividade. Clubes capazes de reduzir seus custos operacionais conseguem praticar preços mais baixos que seus rivais, tornando-se mais interessantes do ponto de vista do torcedor (MATTAR; MATTAR, 2013, p. 105).

Outra fonte de vantagem competitiva do ponto de vista de Mattar e Mattar (2013, p. 106) é relacionada com marketing e ocorre através da imagem da marca junto aos associados, torcedores e consumidores. Para os autores, uma linha de produtos e serviços adequada ao perfil dos torcedores juntamente com uma propaganda e promoções de vendas habilidosas são fontes geradoras de vantagens competitivas nas instituições.

Mattar e Mattar (2013, p. 114) explicitam que para as atividades de marketing progredirem elas necessitam de estruturação, orçamento e estratégias. Os autores ratificam sobre a assertividade das ações de marketing, as quais devem fundamentar-se em dados e informações a respeito do seu ambiente de negócios, do mercado, dos concorrentes e do próprio clube.

Clubes de futebol deveriam adotar sistemas de informações de marketing, conforme reitera Mattar e Mattar (2013, p. 115). O modelo do sistema de informação apresentado pelos autores é formado por quatro subsistemas:

- sistema de monitoração ambiental: consiste em coletar informações relacionadas ao ambiente de negócios da instituição esportiva;
- sistema de informações competitivas: informações públicas a respeito dos concorrentes, obtidas através de reportagens, propagandas, websites e pesquisas;
- sistema de informações internas: manter os gestores informados sobre receitas, despesas, vendas, recursos humanos e desempenho esportivo;
- sistema de marketing: fornece informações sobre o nível de satisfação do torcedor em relação à venda de ingressos, aceitação do torcedor com relação a um novo produto e aceitação com relação ao um novo jogador.

A partir das informações recuperadas através dos subsistemas de informação os gestores de marketing esportivo compreendem o ambiente de negócios e são capazes de tomar decisões.

Do ponto de vista financeiro, clubes de futebol movimentam receitas semelhantes à de grandes corporações (MATTAR; MATTAR, 2013, p. 187). Os autores sustentam este ponto de vista apresentando que no ano de 2010, os clubes participantes da Série A do Campeonato Brasileiro de Futebol investiram US$ 79 milhões apenas na contratação de atletas. Mattar e Mattar (2013, p. 202) sustentam que “o objetivo principal de uma Instituição Esportiva deve ser maximizar a satisfação de seus torcedores e associados”. Os autores concluem que as questões financeiras de clubes de futebol representam o meio de atingir o objetivo principal anteriormente apresentado, a partir de planejamentos financeiros.

A partir dos conceitos salientados nesta seção, a pesquisa vigente irá recuperar os dados do Twitter como forma de auxiliar as áreas de marketing, finanças e planejamento estratégico, de modo a contribuir para o avanço das tomadas de decisão nas esferas apresentadas.

## 2.9 RELAÇÕES ENTRE MINERAÇÃO DE TEXTO E FUTEBOL

Conforme mencionado na seção de justificativa científica, três trabalhos recuperados nas bases de periódicos apresentam estreita relação com esta proposta, conforme detalhamento na sequência.

No trabalho “*Soccer events summarization by using sentiment analysis*”, Aloufi e Saddik (2013) desenvolveram um sistema capaz de analisar sentimentos em tweets de maneira automática coletando os dados através de uma API do Twitter. Após a coleta dos dados, foi realizado o pré-processamento dos dados utilizando algoritmos de aprendizado supervisionado, sendo eles: SVM, Naive Bayes e a rede neural proposta pela ferramenta WEKA. O algoritmo adotado foi o Naïve Bayes, pois apresentou a melhor taxa de acerto. Após a etapa de pré-processamento, selecionou-se as palavras mais relevantes através do método *Term Frequency-Inverse Document Frequency* (TF-IDF). Com isto, os autores classificaram os tweets em cinco classes utilizando os mesmos algoritmos de aprendizado de máquina da etapa de pré-processamento. Para avaliar o programa de classificação os autores adicionaram de

forma manual 1.000 tweets de cada classe, sendo que o algoritmo SVM apresentou o melhor desempenho.

Os autores analisaram os sentimentos a partir de três diferentes perspectivas e métodos. O primeiro método é baseado em um dicionário léxico, onde os autores implementaram tal método a partir da ferramenta R. O segundo método é baseado em um conjunto contendo textos avaliativos cuja linguagem é geralmente subjetiva, sendo esta implementada através do WEKA. O terceiro método é baseado em textos avaliativos, feita a partir da Stanford NLP. O terceiro método foi o que apresentou os melhores resultados, ou seja, foi a perspectiva com o maior número de análise de sentimentos classificados de forma correta.

No artigo de Aloufi e Saddik (2013), “Sentiment Identification in Football-Specific Tweets”, os autores desenvolveram um classificador de sentimentos específico para o futebol, coletando dados através do twitter, especificamente durante a Copa do Mundo FIFA 2014 e a UEFA Champions League 2016/17. Os autores optaram por desenvolver um conjunto de dados de futebol, rotulado manualmente para apoiar a pesquisa na área de análise de sentimentos de futebol. Foram utilizados recursos léxicos como *Bag of Words (BOW), Part of Speech (POS)* e outro desenvolvido para dados específicos de futebol. Feito isso, os autores comparam o desempenho dos algoritmos SVM, Naive Bayes e Random Forrest. O modelo BOW (uni-gram) foi o que obteve a melhor performance dentre os recursos utilizados, enquanto o algoritmo SVM demonstrou ser mais consistente em relação aos outros.

No artigo de Yu e Wang (2015), “World Cup 2014 in the Twitter World: A big data analysis of sentiments in U.S. sports fans tweets”, são analisados tweets referentes a alguns jogos da seleção estadunidense de futebol e outros jogos. Os autores coletaram os dados através de uma API do Twitter e utilizaram a linguagem de programação Python juntamente com a ferramenta *Natural Language Toolkit* (NLTK) disponível em Python para realizar o pré-processamento dos tweets coletados. A análise de sentimentos foi realizada utilizando duas abordagens: uma abordagem léxica escrita na linguagem de programação R e outra abordagem para analisar os *emoticons* (representação gráfica facial utilizando caracteres) e caracterizá-los como positivos ou negativos.

Devido à baixa quantidade de artigos retornados com os termos anteriormente pesquisados, foi realizada uma nova pesquisa, utilizando o termo “Natural Language Processing” aliado com a palavra “Soccer”. A nova busca retornou um total de 37

documentos, dentre os quais somente 1 artigo não era duplicado e tinha relação com o tema.

O documento “Soccer fans sentiment through the eye of Big Data: the UEFA Champions League as a case study” desenvolvido por Aloufi et el. (2018) recuperado pela nova pesquisa aborda a classificação de tweets utilizando o recurso conhecido como *Bag of Words (*BOW) juntamente com os métodos *Term Frequency* (TF) e o *Term Frequency-Inverse Document Frequency* (TF-IDF). A etapa do pré-processamento foi realizada utilizando a ferramenta NLTK. Os autores testaram o BOW utilizando modelos uni-gram e bi-gram, juntamente com recursos léxicos e linguísticos. Além disso, os autores desenvolveram uma análise de sentimento específica para o futebol utilizando as mesmas etapas de pré-processamento anteriormente utilizadas. O algoritmo de aprendizado de máquina utilizado foi o SVM. Ao final do artigo, conclui-se que o modelo uni-gram aliado a característica de BOW e recursos léxicos e linguísticos atingiu o melhor resultado.

As referências teóricas e a bibliografia disponíveis no artigo de Aloufi et al. (2018) foram pesquisadas, em busca da relação com o tema deste projeto de pesquisa. Foram encontrados mais 6 documentos abordando a análise de sentimentos de futebol no Twitter (excluindo os trabalhos já recuperados pela pesquisa na Scopus).

Ao todo, foram encontrados 7 trabalhos relacionados com o tema deste projeto de pesquisa, conforme o quadro 6.

QUADRO 6 - TRABALHOS RECUPERADOS PELA NOVA PESQUISA

| **Título do documento** | **Autores** | **Ano de publicação** |
|---|---|---|
| Soccer fans sentiment through the eye of Big Data: the UEFA Champions League as a case study | Aloufi S;Alzamzami F;Hoda M;El Saddik A | 2018 |
| Goaalll: using sentiment in the World Cup to explore theories of emotion | Lucas Gm;Gratch J;Malandrakis N;Szablowski E;Fessler E;Nichols J | 2017 |
| Opinion mining and sentiment polarity on Twitter and correlation between events and sentiment | Barnaghi P;Ghaffari P;Breslin Jg | 2016 |

| | | |
|---|---|---|
| Soccer events summarization by using sentiment analysis | Jai Andaloussi S;El Mourabit I;Madrane N;Chaouni Sb;Sekkaki A | 2016 |
| World Cup 2014 in the twitter world: a Big Data analysis of sentiments in U.S. sports fans' tweets | Yu Y;Wang X | 2015 |
| A comparison of SVM versus Naive-Bayes techniques for sentiment analysis in tweets: a case study with the 2013 FIFA confederations cup | Alves Alf;De S Baptista C;Firmino Aa;De Oliveira Mg;De Paiva Ac | 2014 |
| Sentiment identification in football-specific tweets | Aloufi S;Saddik Ae | 2013 |

Fonte: O autor (2019)

O texto “Goaalll: Using Sentiment in The World Cup to Explore Theories of Emotion (Lucas et el, 2017) ” Analisa os tweets em inglês da Copa do Mundo FIFA 2014 utilizando o método de aprendizagem de máquina supervisionado Naive Bayes. Os autores utilizaram métodos léxicos do campo conhecido como Parte do Discurso (“*Part of Speech”)* para realizar o pré-processamento da base.

No artigo “*Opinion Mining and Sentiment Polarity on Twitter and Correlation between Events and Sentiment*” os autores Barnaghi, Ghaffari e Breslin (2016) utilizaram o método de aprendizado de máquina baseado em regressão logística bayesiana para classificar os tweets em positivos ou negativos. Foram ainda utilizados outros métodos léxicos, método TF-IDF e Uni-Gram e Bi-Gram.

Por fim, o artigo “*A Comparison of SVM Versus Naive-Bayes Techniques for Sentiment Analysis in Tweets: A Case Study with the 2013 FIFA Confederations Cup*” desenvolvido por Alves et al. (2014) compara dois algoritmos de aprendizado de máquina, sendo o SVM e o Naive Bayes. Os autores analisaram somente tweets escritos em português para realizar a análise de qual algoritmo possui melhor desempenho. Ao final do trabalho, conclui-se que o algoritmo SVM obteve uma melhor performance.

É possível observar que dentre os 7 artigos analisados, 4 deles utilizaram o algoritmo SVM para realizar a análise, sendo o algoritmo com um melhor desempenho geral, segundo os autores.

Feita a discussão da literatura pertinente para este trabalho, a próxima seção apresenta os encaminhamentos metodológicos adotados para o desdobramento da pesquisa.

# 3 ENCAMINHAMENTOS METODOLÓGICOS

Neste capítulo são descritas as trajetórias metodológicas realizadas nesta investigação.

## 3.1 CARACTERIZAÇÃO DA PESQUISA

O presente projeto de pesquisa será realizado na forma de uma pesquisa exploratória, a qual Gil (2002, p. 41) afirma que têm como objetivo proporcionar maior familiaridade com o problema, com vistas a torná-lo mais explícito ou a constituir hipóteses. Segundo o autor, estas pesquisas têm como objetivo principal o aprimoramento de ideias ou a descoberta de intuições, podendo constituir a primeira parte de uma pesquisa mais detalhada.

Quanto à natureza, esta pesquisa é classificada como quantitativa, pois alberga de forma estruturada a coleta e análise de dados quantitativos. A partir dos dados coletados, realiza-se o estudo entre as variáveis, propiciando a interpretação dos resultados.

Com relação aos métodos de investigação, esta pesquisa classifica-se como observacional e experimental, de forma complementar um ao outro. De acordo com Prodanov e Freitas (2013, p. 26) “os métodos gerais oferecem ao pesquisador normas genéricas destinadas a estabelecer uma ruptura entre objetivos científicos e não científicos (ou de senso comum) ”. Gil (2008, p.16) elenca o método experimental através da submissão do objeto de estudo a determinadas variáveis, em condições controladas e conhecidas pelo investigador, a fim de observar os resultados que a variável produz no objeto. O método observacional de acordo com Gil (2008, p. 16) é um dos mais utilizados nas ciências sociais, caracterizado pelo estudo de uma variável sem controlar qualquer ação, apenas observar algo que acontece ou já aconteceu.

A obtenção dos dados coletados nesta pesquisa é conduzida por meio da técnica de observação simples, a qual Gil (2008, p. 101) explica como “aquela em que o pesquisador, permanecendo alheio à comunidade, grupo ou situação que pretende estudar, observa de maneira espontânea os fatos que aí ocorrem”.

O envolvimento do pesquisador nesta pesquisa seguiu o modelo clássico informado por Gil (2008, p. 28), discernido pela absoluta neutralidade em relação ao fenômeno pesquisado, restringindo-se ao que pode ser efetivamente observado.

No desenvolvimento desta pesquisa, as principais etapas são:

- formulação do problema de pesquisa;
- definição dos objetivos gerais e específicos;
- justificativa da pesquisa;
- verificação da literatura existente sobre o tema;
- montagem da base de dados;
- pré-processamento dos dados;
- aplicação de algoritmos de aprendizado de máquina;
- análise e interpretação dos resultados obtidos;
- considerações finais.

Com o propósito de verificar a aderência do tema à realidade das gestões dos clubes envolvidos, foi elaborado um roteiro estruturado no formato de entrevista para ser aplicado nas instituições estudadas nesta pesquisa, (APÊNDICE 1) abordando a utilização do conteúdo veiculado em redes sociais como instrumento de gestão dos clubes. A entrevista foi proposta para verificar questões relacionadas com a mineração de opiniões nas redes sociais dentro dos clubes de futebol, desejando saber se os clubes possuem um departamento relacionado com o tema e como as redes sociais são utilizadas nas tomadas de decisões.

No entanto, após contato com os três clubes pesquisados, apenas o Clube Athlético Paranaense participou da entrevista. Os demais não demonstraram interesse pelo assunto em pauta e não retornaram o contato seja por redes sociais, e-mail, ligações telefônicas ou mensagens. A participação do Clube Athlético Paranaense ocorreu de maneira espontânea e consensual, sendo firmada por meio de um termo de participação (APÊNDICE 2), assinado pelo participante da entrevista.

Durante a entrevista, o clube informou que utiliza quatro redes sociais para interagir com os torcedores, sendo elas: Facebook, Instagram, Twitter e Youtube. No clube existe um departamento responsável pela verificação das postagens e comentários nas redes sociais, desenvolvendo esta função a aproximadamente um ano, mas que essa averiguação é realizada de maneira orgânica, por intermédio de análises pessoais e das métricas disponíveis nas próprias redes sociais. A análise de sentimentos é conhecida no clube e utilizada de modo pontual, a partir das ferramentas disponibilizadas nas redes sociais para analisar as postagens, sem nenhum *software* ou técnica voltada para a mineração de opiniões. Nota-se no clube

um alto número de pessoas integralizando o departamento responsável pelas redes sociais do clube, com aproximadamente 20 pessoas, as quais apresentam formações em distintas áreas, tais como: jornalismo, publicidade, *design* e *marketing.* O clube informou que acredita na utilização das redes sociais como uma forma de contribuição nas tomadas de decisões e que as redes sociais influenciam em alguns aspectos do clube, principalmente nas áreas de *marketing* e relações públicas.

## 3.2 MATERIAIS E MÉTODOS

Para auxiliar na pesquisa, faz-se necessário o uso de algumas ferramentas, como linguagens de programação e softwares com a capacidade de executar os algoritmos de aprendizado de máquina. A primeira ferramenta utilizada foi o Microsoft Excel®, a fim de montar a base de dados e realizar algumas etapas da fase de pré-processamento. Segundo a Microsoft (2019), o Excel permite organizar os dados e padronizá-los em formato de planilhas, permitindo visualizar e descobrir peculiaridades no conjunto de dados.

Dentre as linguagens de programação existentes, as utilizadas serão Python e R, pois, de acordo com uma pesquisa realizada por King e Magoulas (2013), tanto a linguagem Python quanto a linguagem R são preponderantemente utilizadas em problemas de análise de dados. King e Magoulas (2013, p. 7) afirmam que as linguagens apresentadas acima são populares devido a sua facilidade de utilização, código aberto e grande número de bibliotecas específicas para técnicas de aprendizado de máquina. Na linguagem de programação Python, será usado o ambiente *Jupyter* para realizar a criação do código e a aplicação dos métodos. As principais bibliotecas usadas serão: Numpy, *Natural Language ToolKit* (NLTK), Pandas, Tweepy, *SciKit-Learn*, csv, *seaborn* e UnicodeData. Já na linguagem R, a biblioteca utilizada nesta pesquisa foi a Bibliometrix.

A fim de identificar o conteúdo das mensagens, esta pesquisa empregou a aplicação Semantria na sua versão 2016 x64 6.0.181. A ferramenta segundo a Lexalytics (2019) é capaz de analisar textos de 22 idiomas e classificá-los entre positivo, negativo ou neutro. O funcionamento da ferramenta acontece por uma API incorporada na ferramenta Microsoft Excel®. Apesar da ferramenta ser paga, é possível utilizar uma versão corporativa, gratuita por 90 com 8 idiomas inclusos, entre eles o português.

A ferramenta de visualização de dados, *Power BI Desktop,* foi outro recurso empregado neste trabalho. O Power BI é uma solução de análise de negócios que permite visualizar dados e partilhar informações em aplicações ou sites (MICROSOFT, 2019). Nesta pesquisa o *software* serviu para a criação de relatórios de dados e *dashboards.*

Outra aplicação utilizada foi o *software* Orange Canvas, uma ferramenta que segundo Castro e Ferrari (2016, p. 349), é um software gratuito que permite a construção visual, por meio de blocos e fluxogramas, de processos complexos de análise e mineração de dados, sendo baseado na linguagem Python. A ferramenta ainda conta com pacotes adicionais nas áreas de bioinformática, mineração de textos e visualização de dados. O *software* foi utilizado na criação de visualização de dados na forma de nuvem de palavras e em etapas de pré-processamento de texto, conforme será apresentado posteriormente neste documento.

### 3.2.1 Base de dados

A base de dados utilizada foi construída a partir dos tweets coletados da página oficial no Twitter dos Clubes Athlético Paranaense, Coritiba Foot Ball e Paraná Clube.

A criação da base de dados para a aplicação dos algoritmos deu-se a partir de uma aplicação criada em Python para recuperar os tweets. Para isso, foi utilizado o *software* Jupyter Notebook, disponível a partir do Projeto Jupyter, um projeto de código aberto desenvolvido a fim de auxiliar nas aplicações de ciência de dados (JUPYTER, 2019).

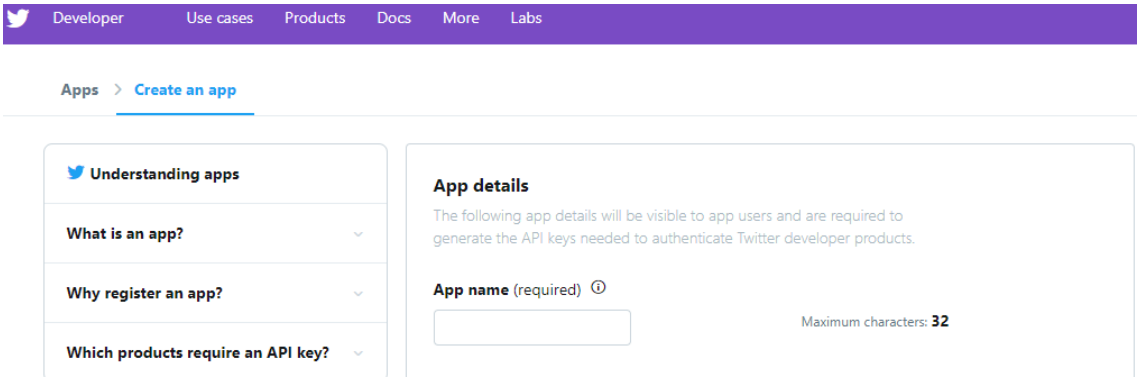
A partir do *Jupyter,* foram utilizadas as bibliotecas oferecidas pela linguagem Python, “Tweepy” e “CSV”, além das expressões regulares, caracterizadas na linguagem Python pela sigle “re”, juntamente com a API disponibilizada pelo Twitter.

A API fornecida pelo Twitter é utilizada por desenvolvedores de aplicações em análise de dados, otimização de anúncios e criação de novas experiências para os usuários da rede social, servindo como um meio de comunicação entre programas de computadores. A aplicação disponibilizada pelo Twitter fornece um vasto acesso aos dados públicos de usuários que optaram por compartilhar suas informações, além do suporte para o gerenciamento das aplicações desenvolvidas (TWITTER, 2019).

A criação da aplicação na rede social é dada acessando o site oficial do Twitter voltado especificamente a desenvolvedores e interessados por criarem aplicações na

rede social, disponível no endereço https://developer.twitter.com/en.html. Após acessar ao site e criar uma conta, é necessário acessar a opção “apps” e clicar em “create an app”, conforme apresenta a figura 8. Para finalizar a criação do aplicativo, é preciso preencher os campos solicitados pela rede social, como o nome, a finalidade e a descrição da aplicação. Com as informações preenchidas, o aplicativo está pronto para ser usado.

FIGURA 8 - CRIAÇÃO DE UMA API NO TWITTER




FONTE: TWITTER (2019).

Com a criação do aplicativo realizada, é preciso informar os dados de autenticação da aplicação, os quais serão utilizados como uma chave de segurança responsável pela conexão entre e a API e o código posteriormente desenvolvido.

Com os passos anteriores realizados, deu-se início a criação da aplicação encarregada por recuperar os dados disponibilizados pelos usuários do Twitter, que servirão como a base de dados deste projeto de pesquisa. A figura 9 detalha a execução da aplicação.

FIGURA 9 - EXECUÇÃO DA APLICAÇÃO RESPONSÁVEL POR RECUPERAR OS TWEETS

```
def search_for_hashtags(consumer_key, consumer_secret, access_token, access_token_secret, hashtag_phrase):

    #create authentication for accessing Twitter
    auth = tweepy.OAuthHandler(consumer_key, consumer_secret)
    auth.set_access_token(access_token, access_token_secret)

    #initialize Tweepy API
    api = tweepy.API(auth)

    #get the name of the spreadsheet we will write to
    fname = '_'.join(re.findall(r"#(\w+)", hashtag_phrase))

    #open the spreadsheet we will write to
    with open('%s.csv' % (fname), 'w', encoding='utf8') as file:

        w = csv.writer(file)

        #write header row to spreadsheet
        w.writerow(['timestamp', 'tweet_text', 'username', 'all_hashtags', 'followers_count'])

        #for each tweet matching our hashtags, write relevant info to the spreadsheet
        for tweet in tweepy.Cursor(api.search, q=hashtag_phrase+' -filter:retweets',tweet_mode='extended').items(1000):
            w.writerow([tweet.created_at, tweet.full_text.replace('\n',' '), tweet.user.screen_name.encode('utf-8'), [e['text
```

FONTE: O autor (2019).

Na criação do código, não foram selecionados tweets em formatos de *retweets*, ou seja, quando uma postagem era compartilhada por outro usuário.

A execução do código retornava um arquivo em formato de valores separados por vírgulas (*Common Separated Values* - CSV) que trazia os seguintes atributos:

- a data em que a postagem foi realizada em formato *timestamp*;
- o conteúdo da postagem;
- o nome de usuário que realizou a postagem;
- as hashtags contidas dentro da postagem;
- e a quantidade de seguidores que o responsável pela postagem possui.

A figura 10 detalha a saída do código criado para recuperar os dados que serão utilizados nas aplicações das técnicas de aprendizado de máquina e mineração de textos.

FIGURA 10 - EXEMPLO DO CONTEÚDO DA BASE DE DADOS

| timestamp | Texto | username | all_hashtags | followers_count |
|---|---|---|---|---|
| 29/09/2019 13:00 | 📸 Se liga nas imagens da vitória sobre o América-MG \| Acesse h | b'Coritiba' | ['MeuCoritiba', 'CFCxAME'] | 947300 |
| 28/09/2019 22:26 | #SerieB \| #CFCxAME \| 2x1 "Fiquei muito entusiasmado com qu | b'Doricoaopobre' | ['SerieB', 'CFCxAME'] | 1630 |
| 28/09/2019 22:21 | #SerieB \| #CFCxAME \| 2x1 Em sua primeira entrevista pós-jogo, | b'Doricoaopobre' | ['SerieB', 'CFCxAME'] | 1630 |
| 28/09/2019 22:10 | #CFCxAME https://t.co/29R5dlxQV0 | b'markviniciosc' | ['CFCxAME'] | 257 |
| 28/09/2019 22:09 | #CFCxAME Estamos na briga. https://t.co/sVOPT2JrwS | b'markviniciosc' | ['CFCxAME'] | 257 |
| 28/09/2019 22:07 | Resultado mais nojento que existe é o tal do 2x1 Ainda bem qu | b'markviniciosc' | ['CFCxAME'] | 257 |
| 28/09/2019 21:41 | #SerieB \| #CFCxAME \| Fim de jogo \| 2 a 1 Coritiba derrota o Am | b'Doricoaopobre' | ['SerieB', 'CFCxAME'] | 1630 |
| 28/09/2019 21:31 | Marcelo de Lima Henrique é o pior árbitro do futebol brasileiro. | b'fabianokbr' | ['CFCxAME'] | 604 |
| 28/09/2019 21:28 | #SerieB \| #CFCxAME \| Fim de jogo \| 2 a 1 A próxima partida do C | b'Doricoaopobre' | ['SerieB', 'CFCxAME'] | 1630 |
| 28/09/2019 21:28 | #SerieB \| #CFCxAME \| Fim de jogo \| 2 a 1 Com o resultado, o Co | b'Doricoaopobre' | ['SerieB', 'CFCxAME'] | 1630 |
| 28/09/2019 21:28 | #SerieB \| #CFCxAME \| Fim de jogo \| 2 a 1 FIM DE JOGO! O Corit | b'Doricoaopobre' | ['SerieB', 'CFCxAME'] | 1630 |
| 28/09/2019 21:27 | FIM DE JOGO: Coritiba 2x1 América-MG #CampeonatoBrasileiro | b'ttPapoReto' | ['CampeonatoBrasileiro', 'SérieB2019', 'CFCxAME'] | 737 |

FONTE: O autor (2019).

Para a recuperação das publicações, definia-se a hashtag[3] a ser recuperada pela aplicação, de forma que foi selecionada as hashtags criadas pelos próprios clubes durante a realização da partida. Optou-se pela recuperação dos dados através das hashtags devido ao fato de ter sido disponibilizada pelo clube como um meio de interagir com os usuários da rede social. Além disso, ao selecionar a hashtag para recuperar as publicações, reduzia-se a possibilidade de ambiguidade com possíveis palavras que pudessem ser semelhantes, mas que não estavam relacionadas com o contexto da pesquisa. A figura 11 apresenta um exemplo da hashtag criada pelo clube como forma de interação.

FIGURA 11 - EXEMPLO DA HASHTAG OFICIAL DISPONIBILIZADA PELO CLUBE


FONTE: Coritiba Foot Ball Club (2019).

A coleta dos dados para a montagem da base ocorreu sempre após o resultado final dos jogos, os quais foram analisados durante o período entre 08/09/2019 e 19/10/2019, com o intuito de coletar dados recentes mas sem que o andamento da partida interferisse na coleta dos dados e consequentemente, no resultado final da análise.

Ao final da coleta dos dados, foram criadas três bases de dados diferentes, de modo que cada base de dados correspondia a um clube especifico. Por depender da quantidade das postagens, as bases possuem números de registros diferentes, de modo que a base do Athlético Paranaense conta com 2435 registros, a do Coritiba com 545 e a do Paraná 435 registros.

[3] De acordo com Boyd, Golder e Lotan (2010), as hashtags (marcadas pelo símbolo "#") são utilizadas como uma forma de rotular os tweets e criar tópicos, possibilitando que outros usuários possam seguir uma conversa centrada em um assunto particular.

A disparidade entre os registros pode residir na diferença entre a quantidade de seguidores que cada clube possui, afetando às interações nas redes sociais. Até o mês de outubro de 2019, o Athlético Paranaense possuía 1.1 milhões de seguidores, o Coritiba 947.6 mil e o Paraná 114,7 mil, evidenciando a grande diferença de registros coletados nas partidas.

Um outro possível fator para uma grande disparidade entre a quantidade de registros coletados reside no aspecto do Athlético Paranaense ter logrado o título de Campeão da Copa do Brasil no mês de setembro de 2019, o que aumentou a popularidade do time e, consequentemente, as interações nas redes sociais durante o período da coleta dos dados. Além disso, o Paraná Clube não utiliza *hashtags* em todas as suas postagens durante o jogo, comprometendo a quantidade de dados coletados.

As coletas dos registros da base do Clube Athlético Paranaense foram realizadas após as partidas realizadas no mês de setembro de 2019, conforme apresenta o quadro 7.

QUADRO 7 - JOGOS DO CLUBE ATHLÉTICO PARANAENSE UTILIZADOS PARA COLETAR OS REGISTROS

| Equipes | Data | Resultado da partida |
|---|---|---|
| Athlético PR x Internacional | 11/09/2019 | 1 x 0 |
| Athlético PR x Avaí | 15/09/2019 | 0 x 1 |
| Athlético PR x Internacional | 18/09/2019 | 2 x 1 |
| Athlético PR x Fortaleza | 26/09/2019 | 4 x 1 |
| Athlético PR x Chapecoense | 29/09/2019 | 1 x 1 |

FONTE: O autor (2019).

Para a montagem da base de dados do Coritiba Football Clube, os jogos utilizados para coletar os dados estão indicados no quadro 8.

QUADRO 8 - JOGOS DO CORITIBA UTILIZADOS PARA COLETAR OS REGISTROS

| Equipes | Data | Resultado da partida |
|---|---|---|
| Coritiba x Atlético Goianiense | 08/09/2019 | 1 x 2 |
| Coritiba x CRB | 21/09/2019 | 0 x 2 |
| Coritiba x América-MG | 28/09/2019 | 2 x 1 |
| Coritiba x Paraná | 05/10/2019 | 0 x 2 |
| Coritiba x Guarani | 08/10/2019 | 1 x 0 |
| Coritiba x Criciúma | 12/10/2019 | 1 x 0 |

| | | |
|---|---|---|
| Coritiba x São Bento | 15/10/2019 | 2 x 1 |

FONTE: O autor (2019).

Finalmente, a montagem da base de dados do Paraná Clube deu-se utilizando os jogos descritos no quadro 9.

QUADRO 9 - JOGOS DO PARANÁ CLUBE UTILIZADOS PARA COLETAR OS REGISTROS

| Equipes | Data | Resultado da partida |
|---|---|---|
| Paraná x Guarani | 21/09/2019 | 0 x 1 |
| Paraná x Ponte Preta | 25/09/2019 | 1 x 1 |
| Paraná x Oeste | 28/09/2019 | 1 x 1 |
| Paraná x Coritiba | 05/10/2019 | 2 x 0 |
| Paraná x Operário | 08/10/2019 | 1 x 0 |
| Paraná x Bragantino | 12/10/2019 | 0 x 2 |
| Paraná x Brasil de Pelotas | 15/10/2019 | 1 x 0 |
| Paraná x Figueirense | 19/10/2019 | 0 x 1 |

FONTE: O autor (2019).

Percebe-se que para a criação da base de da do Athlético Paranaense foi preciso menos jogos em comparação com os demais clubes. Enquanto o Coritiba Football Clube precisou de 7 jogos e o Paraná Clube de 8 jogos, na base de dados do Athlético utilizou-se somente 5 jogos, todos realizados durante o mês de setembro de 2019. Para este trabalho, foi utilizado apenas o conteúdo da mensagem, excluindo os demais atributos. Contudo, para trabalhos futuros poderiam ser utilizados os demais dados, a fim de obter *insights* mais profundos a respeito do perfil dos usuários que realizaram tais postagens.

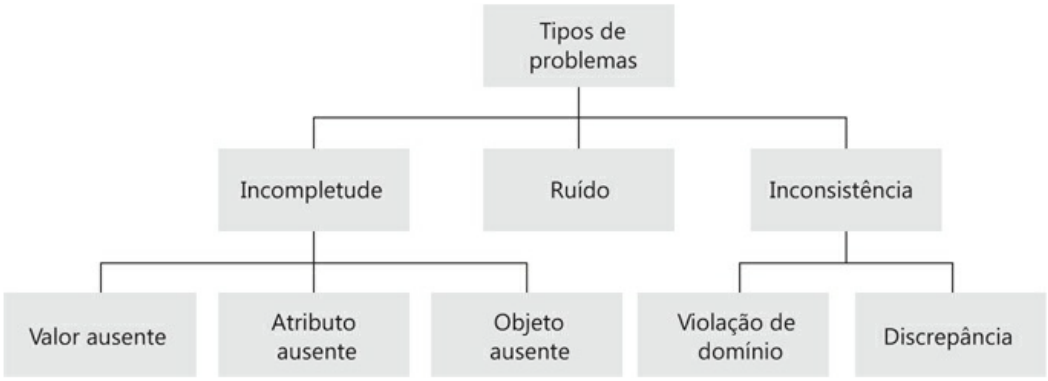
Após a coleta dos dados, iniciou-se as etapas do pré-processamento dos dados, a qual será melhor detalhada no próximo tópico.

### 3.2.2 Pré-processamento

De acordo com Castro e Ferrari (2016, p. 27), a fase de pré-processamento consiste em conhecer e preparar de forma adequada os dados, a fim de tornar o processo de mineração e análise de dados muito mais eficiente e eficaz. Segundo os autores, as etapas de pré-processamento consomem muito tempo e demandam bastante trabalho, sendo um fator determinante para o sucesso do modelo. Os autores ainda elucidam eventuais problemas que podem ocorrer na base de dados, sendo a

etapa de pré-processamento responsável pela eliminação dos problemas, os quais podem ser visualizados na figura 12.

FIGURA 12 - PRINCIPAIS PROBLEMAS COM OS DADOS




FONTE: Castro e Ferrari (2016, p. 27)

Feldman e Sanger (2007, p. 57), afirmam que existem uma grande quantidade de técnicas de pré-processamento no que tange à mineração de textos, com o objetivo principal de estruturar os dados recuperados, preservando suas características. Castro e Ferrari (2016, p. 28) consideram dados estruturados aqueles dados que residem em campos fixos dentro de um arquivo, criados a partir de um modelo de dados, o qual facilita o armazenamento, o acesso e a análise.

Castro e Ferrari (2016, p. 35) elencam as principais tarefas de pré-processamento, sendo elas:

- limpeza: preenchimento de valores ausentes, remoção de ruídos e correção de inconsistências;

- integração: união dos dados coletados através de múltiplas fontes em um único local;

- redução: agrupar ou reduzir atributos redundantes a fim de reduzir a dimensão da base de dados;

- transformação: padronização dos dados para um formato passível de aplicação de diferentes técnicas de mineração;

- discretização: permitir a utilização de métodos que trabalham com atributos nominais em um conjunto maior de problemas.

Para Marcacini, Moura e Rezende (2011, p. 9), um dos maiores desafios do processo de mineração de textos é uma grande dimensionalidade dos dados, de modo que uma coleção de textos pode conter milhares de termos, tornado o processo de extração do conhecimento lento e os resultados menos precisos. A fim de contornar este problema, utiliza-se um subconjunto conciso e representativo de termos da coleção total, através da seleção de termos.

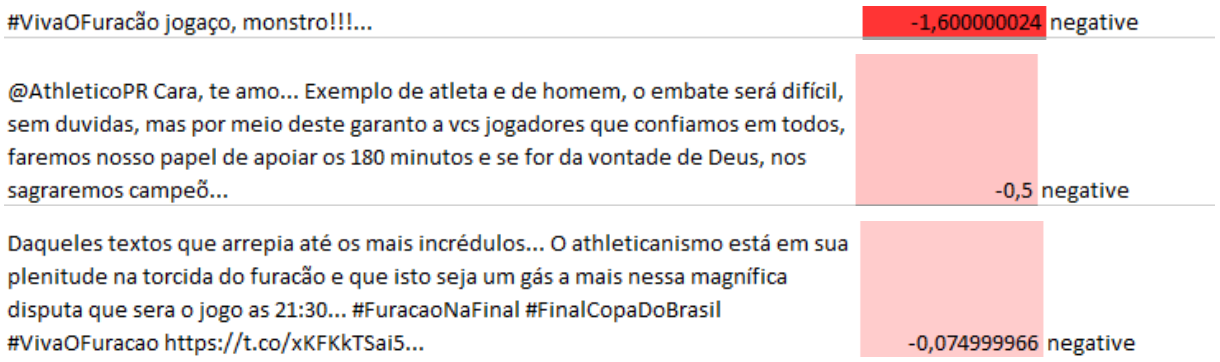
Na visão de Marcacini, Moura e Rezende (2011, p. 9), o primeiro passo na seleção de termos é a eliminação de *stopwords,* que são termos sem grande representatividade, tais como artigos, pronomes e advérbios. O segundo passo consiste na identificação das variações morfológicas e termos sinônimos através de processos de *stemming*, os quais consistem na redução das palavras à sua raiz. Também é possível selecionar os termos através de medidas estatísticas simples, como a frequência que cada termo aparece em um conjunto de documentos.

Neste estudo, foram adotadas as seguintes ferramentas de pré-processamento:

- **Microsoft Excel®:** a ferramenta foi utilizada para armazenar os dados e fazer a seleção das colunas a serem utilizadas, além da redução da quantidade de registros, eliminando aqueles que não continham nenhuma informação (linhas em branco). As colunas “data”, “nome de usuário”, “hashtags” e “quantidade de seguidores” foram deletadas nesta etapa, deixando somente a coluna “texto”, a qual continha a publicação realizada no Twitter. Após a exclusão das colunas, uma nova coluna foi adicionada a base de dados. A coluna “Polaridade” foi criada com o intuito de rotular os dados, informando se o conteúdo da coluna “texto” apresentava uma mensagem positiva, negativa ou neutra. A rotulação ocorreu de forma manual, de modo que cada tweet foi visualizado e a partir do conteúdo da postagem recebeu um rótulo com o sentimento contido neste tweet. Optou-se por utilizar a rotulação manual após verificar que ferramentas de sentimentalização apresentavam dificuldades para rotular os dados de maneira adequada. Como pode ser visto na figura 13, a ferramenta de análise de texto e sentimentos, Semantria, classificava alguns dados de modo incorreto. Na figura 13, as instâncias foram definidas como negativas, quando na verdade apresentam sentimentos positivos, mas que utilizam palavras como

“monstro”, “difícil”, “arrepia” e “incrédulos”, induzindo a ferramenta para uma classificação errada. Dentre os possíveis aspectos para tais imprecisões, destacam-se a dificuldade de as ferramentas descobrirem ironias e aspectos linguísticos, além da falta de contexto com relação aos dados analisados.

FIGURA 13 - EXEMPLOS DE CLASSIFICAÇÕES INCORRETAS FEITAS PELA FERRAMENTA SEMANTRIA



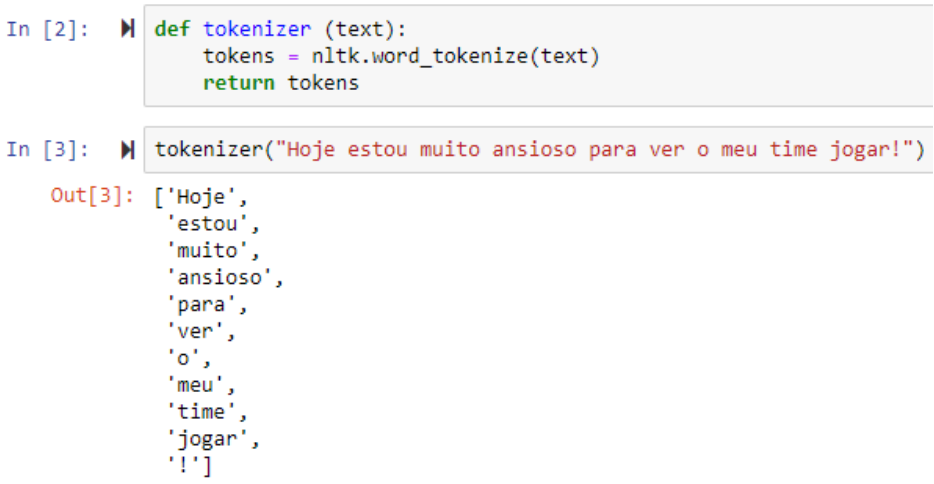

FONTE: O autor (2019).

Outra etapa que ocorreu no Microsoft Excel® foi a seleção aleatória de 500 registros, os quais serão incorporados nas demais etapas do fluxograma e ao final, serão utilizados pelos algoritmos de aprendizado de máquina para classificar o sentimento contido nos dados. Para selecionar aleatoriamente os dados, foi utilizada a fórmula “ALEATÓRIOENTRE”, a qual através de dois valores utilizados como parâmetros, gera números aleatórios a cada nova ação realizada dentro da planilha. Com os números randômicos gerados, criou-se um índice, de modo que cada número representava uma linha contendo uma postagem da base de dados. Com a ordenação dos valores do índice, utilizou-se a fórmula “PROCV”, de maneira onde cada número encontrado retornava a célula contendo a postagem e a polaridade que era representada pelo número aleatório. Os 500 primeiros registros foram selecionados e inseridos em uma nova planilha vazia, somente com a postagem e a polaridade.

- **Python**: algumas etapas de pré-processamento foram programadas nesta linguagem, incluindo a padronização os dados. De acordo com Castro e Ferrari (2016, p. 50), o objetivo da padronização é resolver

problemas de conformidade entre os dados, a fim de facilitar aplicações de algoritmos de mineração. Para realizar a padronização dos dados, o código desenvolvido trabalha com as etapas de tokenização, *stemming*, remoção de caracteres especiais como *emoticons* e pontuações, remoção de *links,* eliminação de *stopwords* e padronização para todas as palavras ficarem em letras minúsculas, recurso este conhecido como capitalização. A etapa de tokenização segundo Clark, Fox e Lappin (2010, p. 534) consiste em segmentar o texto em suas unidades lexicais, delimitando as palavras através dos espaços em branco. Para a realização desta etapa, utilizou-se o ferramental de linguagem natural disponível em Python, conhecido como NLTK, conforme apresenta a figura 14.

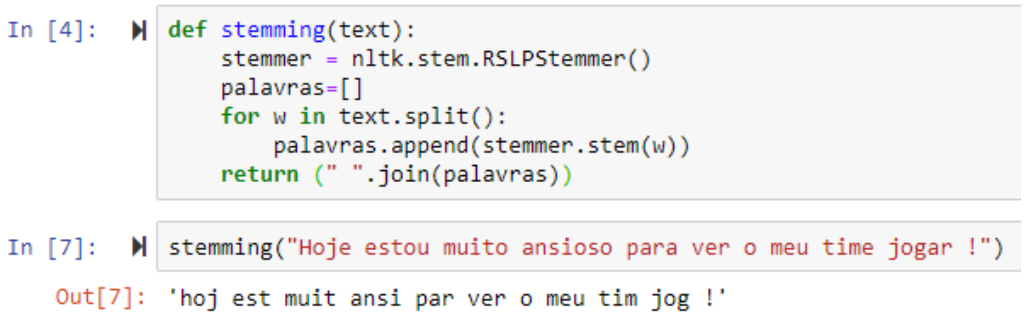
FIGURA 14 - CÓDIGO RESPONSÁVEL PELA TOKENIZAÇÃO DAS PALAVRAS




FONTE: O autor (2019).

A próxima etapa do pré-processamento foi o *stemming*, caracterizado por Jivani (2011, p. 1930) pela manipulação das terminações das palavras, reduzindo-as às suas raízes e eliminando quaisquer sufixos e prefixos anexados ao termo, conforme indicado pela figura 15. Nesta etapa utilizou outra biblioteca disponível em Python, conhecida como “RSLPStemmer”, responsável por realizar o *stemming* das palavras em português.

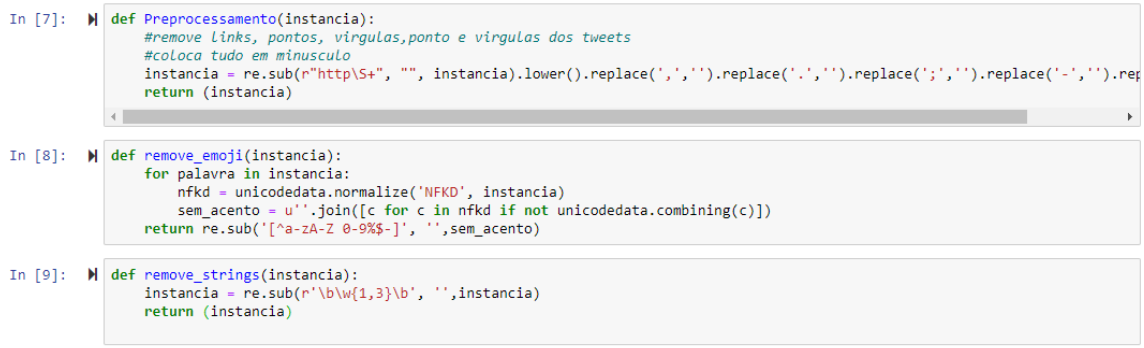
FIGURA 15 - EXECUÇÃO DO CÓDIGO RESPONSÁVEL PELO STEMMING




FONTE: O autor (2019).

Após isto, foi necessário eliminar caracteres especiais, como pontuações, sinais e emoticons a fim de eliminar os ruídos e melhorar a execução do algoritmo de aprendizado. Também foi criado um código responsável por eliminar palavras com menos de três caracteres, uma vez que tais palavras não são capazes de formar palavras, e consequentemente, não servem para realizar a análise. Os códigos baseiam-se nas expressões regulares, as quais já foram utilizadas anteriormente. A figura 16 detalha a aplicação do código.

FIGURA 16 - EXECUÇÃO DO CÓDIGO RESPONSÁVEL POR ELIMINAR CARACTERES ESPECIAIS

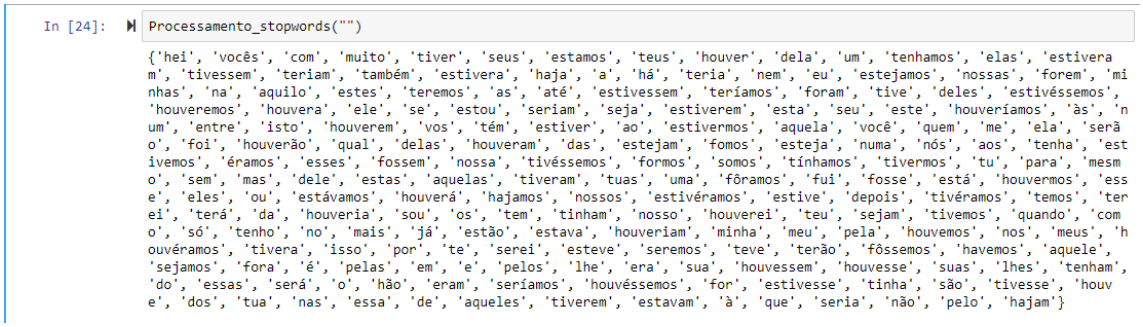



FONTE: O autor (2019).

Outra etapa do pré-processamento dos dados consistiu em remover *stopwords,* conceituadas por Saif, Fernandez e Alani (2014, p. 282) como sendo palavras na análise de sentimentos com um fraco contexto semântico e incapazes de expressar sentimentos. Sendo assim, as

remoções destas palavras contribuem para uma análise mais assertiva. Para esta etapa, novamente foi utilizada a biblioteca “NLTK”, que conta com um pacote em português com palavras consideradas *stopwords.* A figura 17 indica quais palavras são consideradas *stopwords*.

FIGURA 17 - PALAVRAS CONSIDERADAS STOPWORDS NA BIBLIOTECA NLTK




FONTE: *Natural Language ToolKit* (2019).

Contudo, com o intuito de aprimorar ainda mais a análise, foi criado um código para atualizar a lista de palavras consideradas *stopwords* descritas na figura 15. Este código retira palavras listadas pela biblioteca, como a palavra “não”, que para a análise de sentimentos neste trabalho possui alta importância, uma vez que a palavra é muitas vezes responsável por alterar o sentimento de uma frase. Além da exclusão de palavras, novas palavras foram consideradas *stopwords*, visto que não apresentavam forte contexto semântico. Na figura 18 está exposto o código desenvolvido para atualizar a lista de palavras consideradas *stopwords* e realizar a eliminação das mesmas. Palavras como “vai”, “ter”, “quando”, “de”, “sua” e diversas outras foram consideradas fracas semanticamente e desta forma, foram consideradas *stoptowords*.

FIGURA 18 - CÓDIGO PARA REMOVER STOPWORDS

```
In [18]:
def PreprocessamentoSemStopWords(instancia):
    #remove links dos tweets
    instancia = re.sub(r"http\S+", "", instancia).lower().replace(',','').replace('.','').replace(';','').replace('-','').rep

    #remove stopwords
    stopwords = set(nltk.corpus.stopwords.words('portuguese'))


    #adicionar novas palavras
    new_stopwords = ['vai', 'ter', 'quando', 'isso', 'pro', 'pra', 'ta', 'ai', 'de','mais','seu','sua','era','seja','qual', '
    new_stopwords_list = stopwords.union(new_stopwords)

    #remover palavras da lista de stopwords
    not_stopwords = {'não','nao'}
    final_stopwords = set([word for word in new_stopwords_list if word not in not_stopwords])

    palavras = [i for i in instancia.split() if not i in final_stopwords]
    return (" ".join(palavras))
```

FONTE: O autor (2019)

Após as etapas de pré-processamento, os textos contidos no tweets estão prontos para serem utilizados na aplicação dos algoritmos, uma vez que estão normalizados.

As figuras 19 e 20 exemplificam a diferença dos termos mais frequentes de uma mesma base, mas uma sem as etapas de pré-processamento e outra após executar o código responsável pela normalização da base de dados. A criação da nuvem de palavras foi feita na ferramenta Orange, utilizando a base de dados criada para o clube Athlético Paranaense como exemplo.

FIGURA 19 - NUVEM DE PALAVRAS COM OS TERMOS MAIS FREQUENTES DA BASE SEM PRÉ-PROCESSAMENTO


FONTE: O autor (2019).

A partir das imagens 19 e 20, é possível notar a importância da etapa de pré-processamento dos dados, uma vez que termos dispensáveis e sem contribuição para a análise dos dados foram eliminados, como símbolos, caracteres especiais, números e palavras com fraco contexto semântico. Apesar da etapa de pré-processamento ter sido realizada, ainda se nota falhas no pré-processamento, como por exemplo, as palavras “capxf”, “capxch” e “capx”, as quais eram *hashtags* criadas pelo próprio Athlético Paranaense, utilizando a abreviação do nome do clube (Clube Athlético Paranaense), aliada à letra “X” juntamente com a abreviação do time adversário. Nestes casos, a sigla do clube adversário foi removida no pré-processamento, prejudicando a nuvem de palavras e formando termos incompreensíveis. Apesar disto, o pré-processamento serviu para agrupar palavras semelhantes, uma vez que a etapa de *stemming* elimina os radicais das palavras. Termos como, “chutar”, “chutou”, “chute” e “chutando” foram todos agrupados para o termo “chut”, apresentando os termos mais relevantes.

FIGURA 20 - NUVEM DE PALAVRAS COM OS TERMOS MAIS FREQUENTES DA BASE APÓS PRÉ-PROCESSAMENTO


FONTE: O Autor (2019).

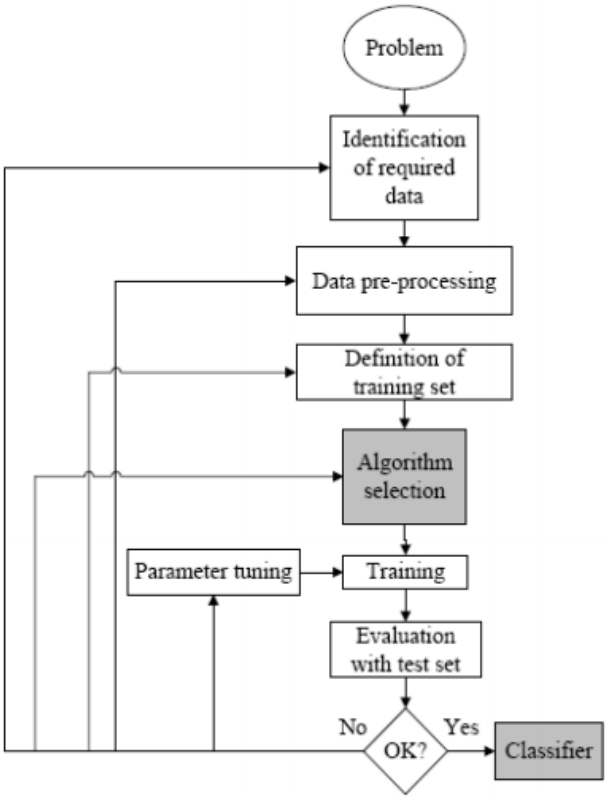
Foi criado um arquivo no código de programação responsável por armazenar as saídas dos textos pré-processados em formato CSV. Este arquivo foi utilizado para a etapa de aplicação dos algoritmos, detalhada na próxima seção desta pesquisa.

### 3.2.3 Processamento

Na etapa de processamento dos dados, utilizou-se a linguagem de programação Python para a aplicação dos algoritmos de classificação, uma vez que segundo McKinney (2018, p. 2), o Python desenvolveu uma grande comunidade na área de ciência da computação e análise de dados, sendo uma das mais importantes linguagens para ciência de dados, aprendizado de máquina e desenvolvimento de aplicações. McKinney (2018, p. 2) sustenta que o suporte da linguagem de programação para as bibliotecas voltadas para a ciência de dados tornou-a extremamente popular na aplicação de tarefas de análise de dados.

Utilizando-se da linguagem de programação Python no ambiente *Jupyter* para o processamento dos dados, serão utilizados algoritmos de aprendizado de máquina supervisionados. Para Osisanwo et al.. (2017, p. 129), a classificação de dados utilizando técnicas de aprendizado de máquina supervisionado ocorre após as etapas de identificação do problema, coleta dos dados necessários, pré-processamento dos dados, definição do conjunto de treinamento e seleção do algoritmo, de tal modo que o aprendizado do classificador é dado a partir do conjunto de instâncias de treinamento, criando um classificador que pode ser utilizado para generalizar os dados a partir de novas instâncias, conforme indicado na figura 21. Constata-se no modelo de Osisanwo et al.. (2017) que caso o modelo não atenda a solução do problema de maneira satisfatória, uma verificação de todas as etapas anteriores, com o intuito de melhorar o processo de aprendizado de máquina supervisionado.

FIGURA 21 - PROCESSO DE APRENDIZADO DE MÁQUINA SUPERVISIONADO




FONTE: Osisanwo et al. (2017)

A respeito da etapa de seleção dos algoritmos descrita por Osisanwo et al. (2017, p. 129), os algoritmos escolhidos para a classificação dos dados foram o Multinomial Naive Bayes[4], o SVM[5] e um modelo de árvore de decisão (CART[6]) disponível na biblioteca Python “Scikit-Learn”[7]. Tais algoritmos de classificação são apresentados por Osisanwo et al. (2017, p. 129) como alguns utilizados no aprendizado de máquina supervisionado.

A escolha para tais algoritmos deu-se na etapa de investigação das pesquisas já desenvolvidas na área de análise de sentimento, uma vez que os algoritmos selecionados foram amplamente utilizados por outros pesquisadores em trabalhos

[4] MCCALLUM, Andrew; NIGAN, Kamal. A Comparison of Event Models for Naive Bayes Text Classification. **Work Learn Text Categ**, Pittsburgh, v. 752, n. 1, maio 2001.
[5] VALENTINI, Giorgio; DIETTERICH, Thomas G. Bias-variance analysis of support vector machines for the development of SVM-based ensemble methods. **Journal of Machine Learning Research**, New York, v. 5, n. 1, p.725-775, jul. 2004
[6] BREIMAN, Leo et al.. **Classification and Regression Trees.** Belmont: Wadsworth International Group, 1984.
[7] Pedregosa, F et al.. Scikit-learn: Machine Learning in {P}ython. Journal of Machine Learning Research, Si, v. 212, n. 1, p.2825-2830, 2011

anteriores, como Aloufi e Saddik (2013), Alves et al. (2014), Lucas et al. (2017) e Aloufi et al. (2018).

Para dar início a etapa de processamento, foi realizada a instalação das bibliotecas de aprendizado de máquina necessárias, conforme indica a figura 22.

FIGURA 22 - INSTALAÇÃO DAS BIBLIOTECAS DE APRENDIZADO DE MÁQUINA

```
import pandas as pd
import numpy as np
from nltk.tokenize import word_tokenize
from nltk import pos_tag
from nltk.corpus import stopwords
from nltk.stem import WordNetLemmatizer
from sklearn.preprocessing import LabelEncoder
from collections import defaultdict
from nltk.corpus import wordnet as wn
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn import model_selection, naive_bayes, svm
from sklearn.model_selection import cross_val_predict
from sklearn.metrics import accuracy_score
from sklearn import metrics
```

FONTE: O autor (2019).

As bibliotecas utilizadas foram “pandas”, “numpy”, “nltk” e “sklearn”, todas utilizadas para realizar ações focadas em aprendizado de máquina e análise de dados.

Com a instalação das bibliotecas realizadas, os 500 registros aleatórios anteriormente selecionados com a utilização da ferramenta Microsoft Excel® foram incorporados ao código Python. A base de dados foi inserida no código e dividida em duas classes, uma contendo os textos e que será chamada de *corpus* e outra contendo as polaridades de cada registro. Após a divisão da base de dados em duas classes, foi realizada a etapa de preparação do *corpus* para as fases de treinamento e de testes. Segundo Castro e Ferrari (2016, p. 156), a etapa de separação dos dados em conjuntos de treinamento e testes faz parte do processo de predição dos dados, o qual consiste em automatizar os processos de tomada de decisão. A figura 24 detalha os passos na projeção do modelo preditivo segundo os autores.

FIGURA 23 - CONSTRUÇÃO E APLICAÇÃO DE UM MODELO PREDITIVO


FONTE: Castro e Ferrari (2016, p. 157)

A etapa de pré-processamento foi descrita na seção anterior deste projeto de pesquisa e com isso, não será detalhada nesta seção. Para Castro e Ferrari (2016, p. 156), a etapa de separação dos dados em um conjunto de treinamento serve para gerar o modelo preditivo, enquanto a parte de testes é utilizada para avaliar a qualidade do modelo gerado. Deste modo, o *corpus* deste trabalho foi definido na proporção de 70% dos dados para treinamento e 30% para os testes e avaliação do modelo.

Para aprimorar a leitura dos dados pelos modelos de aprendizado de máquina, os dados em formato de texto foram transformados para valores numéricos e posteriormente, vetorizados através do método TF-IDF (*Term Frequency – Inverse Document Frequency)*, o qual serve para ponderar a frequência dos termos dentro do *corpus*. Conforme apresenta Robertson (2004, p. 503), o método TF-IDF surgiu após a criação de outras duas medidas, sendo elas:

- IDF (*Inverse Document Frequency):* este método baseia-se na ideia de que um termo muito frequente não é um bom discriminador e deve conter um peso menor do que termos que aparecem com menos frequência;
- TF (*Term Frequecy):* este método verifica a frequência com que as palavras aparecem nos documentos.

Desta forma, o método TF-IDF envolve a multiplicação da medida IDF pela medida TF com o intuito de apresentar as palavras mais importantes no conjunto de documentos. Neste projeto de pesquisa, definiu-se parametrizar as 50 palavras mais frequentes do *corpus* e apresenta-las em valores numéricos. A figura 24 apresenta as

etapas de separação dos dados nos conjuntos de treinamento e testes, transformação dos textos para valores numéricos e aplicação do método TF-IDF.

FIGURA 24 - SEPARAÇÃO DOS DADOS, TRANSFORMAÇÃO PARA NÚMEROS E APLICAÇÃO DO MÉTODO TF-IDF

```
#Separação dos dados em conjunto de treinamento e teste

Train_X, Test_X, Train_Y, Test_Y = model_selection.train_test_split(df['texto'], df['Polaridade'], test_size = 0.2)
```

```
#Transformação dos textos para valores numéricos
Encoder = LabelEncoder()
Train_Y = Encoder.fit_transform(Train_Y)
Test_Y = Encoder.fit_transform(Test_Y)
```

```
#Aplicação do método tf-idf com as 5000 palavras mais interessantes

Tfidf_vect = TfidfVectorizer(max_features = 5000)
Tfidf_vect.fit(df['texto'])

Train_X_Tfidf = Tfidf_vect.transform(Train_X)
Test_X_Tfidf = Tfidf_vect.transform(Test_X)
```

FONTE: O autor (2019).

Após as etapas de pré-processamento, os dados estão prontos para serem aplicados nos algoritmos de classificação. O processamento dos dados ocorreu utilizando os algoritmos de classificação já mencionados, Naive Bayes, SVM e árvore de decisão (CART), devido às suas aplicabilidades no campo da análise de sentimentos e classificação de dados.

Primeiramente, foi necessário definir as variáveis de treino que o modelo de classificação iria interpretar. Após o treinamento do algoritmo, selecionou-se os dados de teste para o modelo realizar a predição dos dados. A figura 25 apresenta as variáveis de treino do modelo e a predição do algoritmo.

FIGURA 25 - APLICAÇÃO DO ALGORITMO NAIVE BAYES NOS DADOS DE TESTE

```
##### Multinomial Naive Bayes #####

#fit the training dataset on the NB classifier

Naive = naive_bayes.MultinomialNB()
Naive.fit(Train_X_Tfidf,Train_Y)

#predict the labels on dataset

predictions_NB = Naive.predict(Test_X_Tfidf)
```

FONTE: O autor (2019).

A partir da predição da variável de teste, é realizada a verificação das métricas de classificação do modelo, buscando avaliar a assertividade do modelo.

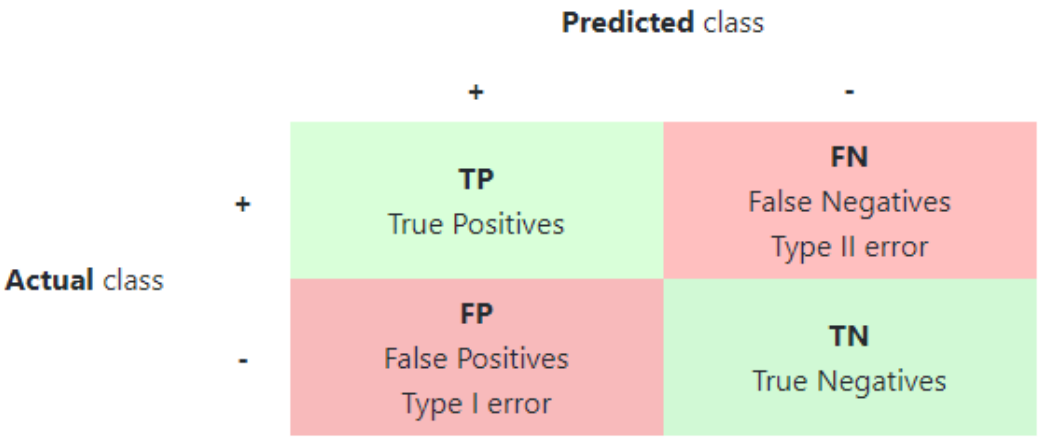
### 3.2.4 Pós-processamento

Esta etapa visa apresentar as formas de avaliação dos resultados obtidos com os algoritmos de classificação apresentados na etapa de processamento.

Uma das formas de avaliação do modelo indicada por Amidi e Amidi (2019) é a matriz de confusão, utilizada para avaliar a assertividade geral do modelo. A matriz de confusão apresenta o número de instâncias reais e o número de instâncias preditas, sendo um indicador do número de classificações corretas e incorretas em cada classe.

A matriz de confusão apresenta, de acordo com Armah, Luo e Quin (2014), quatro resultados: verdadeiros positivos (*true positives* - TP), falsos positivos (*false positives* - FP), verdadeiros negativos (*true negatives* - TN) e falsos negativos (*false negatives* - FN). TP são casos em que o modelo classificou um valor corretamente como positivo, enquanto TN são valores negativos classificados corretamente pelo modelo, ou seja, TP e TN são resultados corretamente preditos pelo modelo. Já valores FP e FN são caracterizados por serem erroneamente classificados como positivos e negativos, respectivamente. Desta forma, pode-se avaliar a assertividade de um modelo analisando a diagonal principal da matriz de confusão, conforme indicado pelos valores destacados em verde na figura 26.

FIGURA 26 - MATRIZ DE CONFUSÃO DE UM ALGORITMO DE CLASSIFICAÇÃO




FONTE: Amidi e Amidi (2019).

A matriz de confusão e seus resultados possibilitam novas métricas de avaliação do modelo, sendo elas:

- Acurácia: métrica utilizada para medir a performance geral do modelo, dada pela fórmula:

$$A = \frac{TP + TN}{TP + TN + FP + FN}$$

- Precisão: o quão preciso o modelo previu os valores positivos, caracterizada pela fórmula:

$$P = \frac{TP}{TP + FP}$$

- Recall: capacidade do modelo para encontrar exemplos corretamente classificados:

$$R = \frac{TP}{TP + FN}$$

- F1 Score: combinação da métricas precisão e recall, utilizada para avaliar conjuntos de dados desproporcionais do modelo. Seu cálculo é dado pela seguinte fórmula:

$$F1 = \frac{2TP}{2TP + FP + FN}$$

A partir das métricas destacadas anteriormente, criou-se um código capaz de apresentar os resultados obtidos, conforme ilustrado na figura 27.

FIGURA 27 - CÓDIGO PARA AVALIAR O MODELO DE CLASSIFICAÇÃO

```
print("Decision Tree Score = ", accuracy_score(Tree, Test_Y)*100)

target_names = ['Positivo', 'Neutro', 'Negativo']
print(metrics.classification_report(Tree, Test_Y, target_names = target_names, digits = 3))
```

FONTE: O autor (2019).

A saída do código desenvolvido na figura 27 com os resultados obtidos é dada na forma de uma matriz, a qual apresenta os valores para análise da classificação dos valores. A figura 28 demonstra a saída da avaliação do modelo com as métricas destacadas anteriormente.

FIGURA 28 - RESULTADOS DO MODELO DE CLASSIFICAÇÃO

```
Decision Tree Score =  65.70841889117042
              precision    recall  f1-score   support

    Positivo      0.164     0.290     0.209        31
      Neutro      0.849     0.725     0.782       349
    Negativo      0.433     0.542     0.481       107

    accuracy                          0.657       487
   macro avg      0.482     0.519     0.491       487
weighted avg      0.714     0.657     0.680       487
```

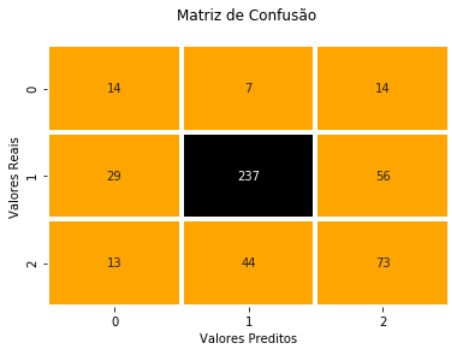
FONTE: O autor (2019).

A imagem 28 retorna os valores de precisão, *recall* e *F1-score* de cada rótulo do modelo, apresentando estes valores na forma de uma matriz. Além disso, a saída do código demonstra a acurácia geral do modelo e o número de instâncias que foram utilizadas para realizar o teste do modelo preditivo.

Desenvolveu-se outro código a fim de apresentar a matriz de confusão, utilizando outras bibliotecas disponíveis na linguagem Python, sendo elas, “seaborn” e “matplotlib”. As bibliotecas são utilizadas para apresentar de maneira visual a quantidade de valores reais e valores preditos do modelo. A figura 29 demonstra o código desenvolvido e a matriz de confusão com os valores reais e preditos.

FIGURA 29 - APLICAÇÃO DA MATRIZ DE CONFUSÃO DO MODELO

```
import seaborn as sns
import matplotlib.pyplot as plt
sns.heatmap(confusion_matrix(Tree,Test_Y), annot=True, lw=3,cbar=False,linewidths=.5,fmt="d", cmap = ['orange','black'])
plt.ylabel("Valores Reais")
plt.xlabel("Valores Preditos")
plt.title("Matriz de Confusão \n")
plt.show()
```




FONTE: O autor (2019).

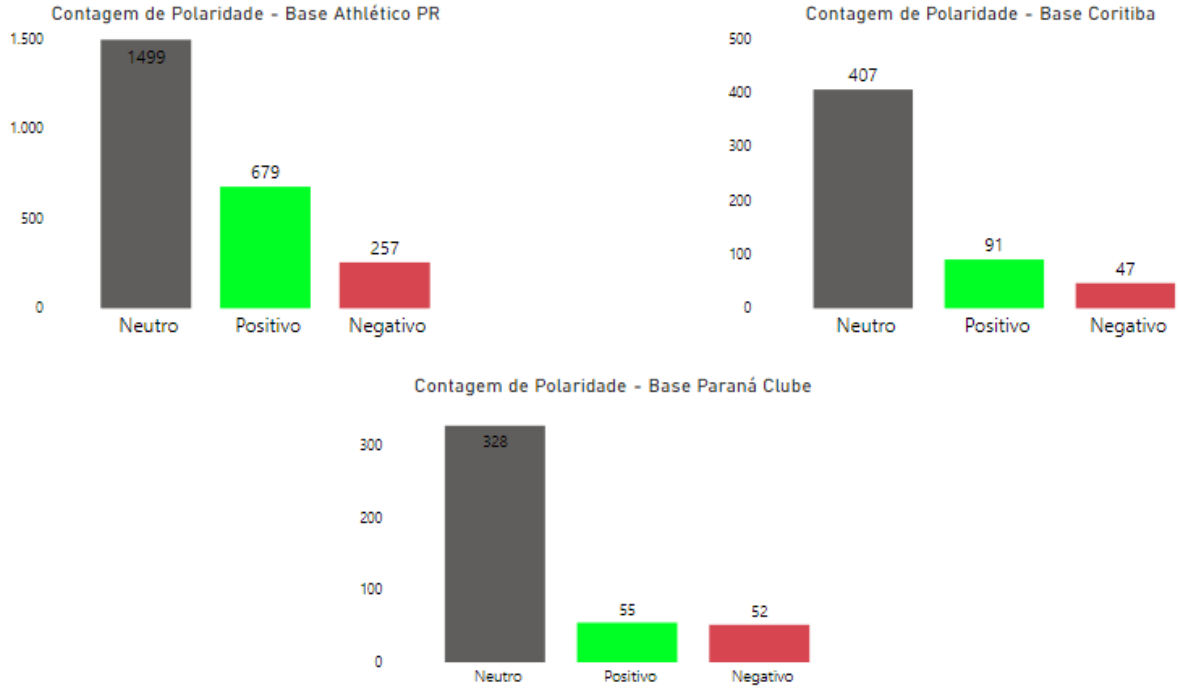
# 4 RESULTADOS

Nesta seção são evidenciados os resultados atingidos com a aplicação dos algoritmos de aprendizagem de máquina nas bases de dados pré-processadas.

## 4.1 ANÁLISE DAS BASES DE DADOS

Com a rotulação dos dados realizada de maneira manual, buscou-se verificar a quantidade de registros existentes de cada tipo (positivo, negativo ou neutro), obtendo maior profundidade com os dados. A figura 30 demonstra a quantidade de registros positivos, negativos e neutros nas três bases criadas.

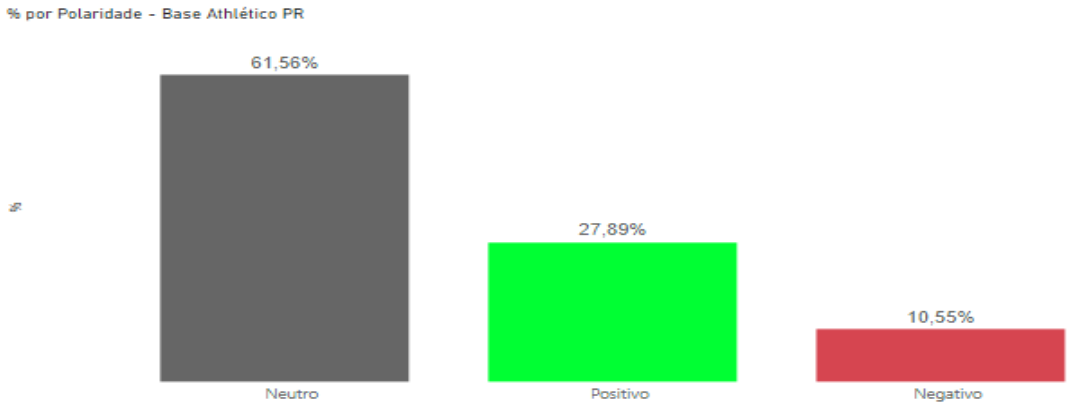
FIGURA 30 - QUANTIDADE DE REGISTRO POR POLARIDADE DAS BASES DE DADOS




FONTE: O autor (2019.

A partir dos gráficos gerados, nota-se a grande quantidade de registros neutros nas bases de dados, ou seja, registros que não demonstram caráter sentimental. Demonstrado estes valores em porcentagens, é possível verificar com ainda mais precisão a disparidade entre a quantidade de rótulos neutros de rótulos positivos e negativos. As imagens 31, 32 e 33 apresentam a porcentagem de cada polaridade contida na base do Athlético Paranaense, Coritiba e Paraná Clube, respectivamente.

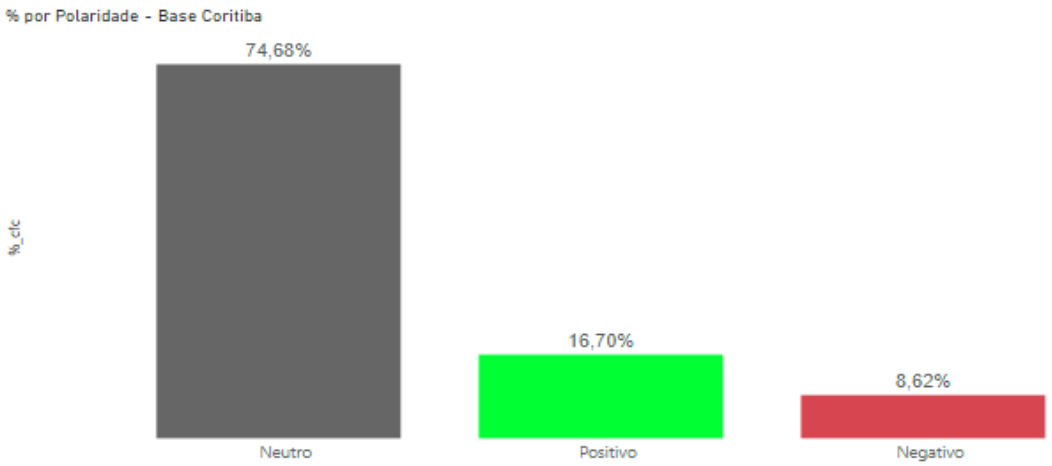
FIGURA 31 - VALORES EM PORCENTAGEM DAS POLARIDADES DA BASE DO ATHLÉTICO PR




FONTE: O autor (2019).

Dentre as bases, a do Athlético Paranaense é a com o menor número de instâncias neutras e a maior quantidade de valores positivos. Dentro os possíveis causadores para a alta quantidade de registros positivos, a competição vencida pelo clube, já mencionada anteriormente, pode ter sido fundamental para o aumento deste tipo de registros. Além disto, nota-se poucos valores negativos na base, sendo a base de dados com a maior diferença entre a porcentagem de registros positivos dos registros negativos.

FIGURA 32 - VALORES EM PORCENTAGEM DAS POLARIDADES DA BASE DO CORITIBA

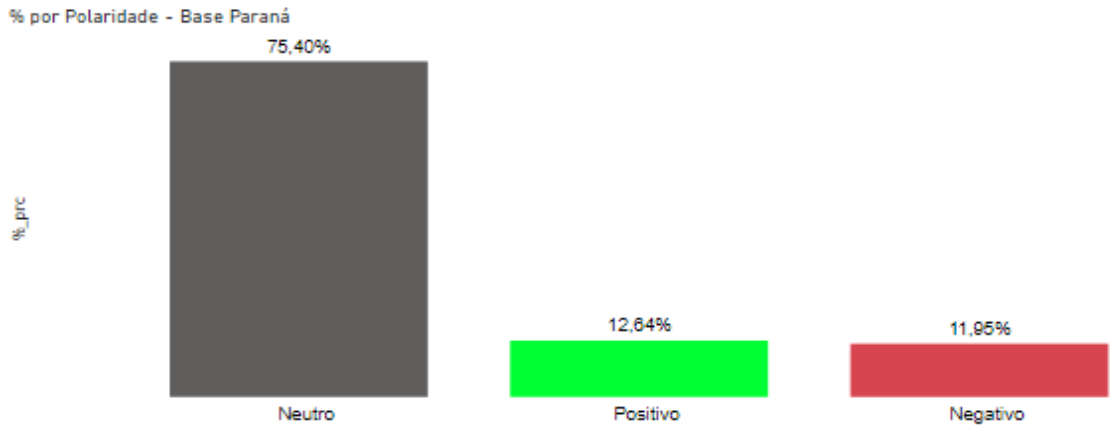



FONTE: O autor (2019).

A base de dados do Coritiba se destaca por apresentar a menor porcentagem de registros negativos quando comparadas com as bases de Athlético Paranaense e

Paraná Clube. A quantidade de registros do tipo “positivo” possui quase o dobro de registros do tipo negativo, demonstrando que apesar da baixa quantidade de registros com polaridades negativas ou positivas, as de caráter positivo são predominantes.

FIGURA 33 - VALORES EM PORCENTAGEM DAS POLARIDADES DA BASE DO PARANÁ CLUBE




FONTE: O autor (2019).

A base do Paraná Clube é caracterizada por ser a com a maior porcentagem de registros neutros, além de ser a mais equilibrada entre a quantidade de registros positivos e negativos. Um alto equilíbrio nos dados denota que os torcedores apresentam divergências com relação ao sentimento para com o time, de modo que não existe um consenso com relação ao desempenho do clube.

As imagens demonstram que todas as bases possuem mais da metade de seus registros categorizados como neutros, causando maiores dificuldades para o modelo classificar corretamente as demais instâncias, uma vez que existem menos dados de treinamento e testes.

Com base nos tweets coletados, outra análise realizada consistiu em verificar os termos mais frequentes dos dados positivos e negativos de cada base, demonstrando as palavras mais comuns responsáveis por classificar uma instância como positiva ou negativa por meio de uma nuvem de palavras. Para isso, utilizou-se as bases de dados originais, sem que os dados tivessem qualquer tipo de pré-processamento ou manipulação juntamente com a ferramenta *Orange* em seu módulo de mineração de textos. Após inserir os dados positivos ou negativos na ferramenta, foi realizado as etapas de pré-processamento, alterando todos os dados para letras minúsculas (capitalização), removendo acentos, caracteres especiais, *links* e

*stopwords*, além do processo de tokenização. As *stopwords* foram definidas conforme uma lista de termos disponível no apêndice 3. Optou-se por realizar um novo pré-processamento dos dados na ferramenta devida a facilidade de manipulação dos processos de normalização dos dados e a retirada da função de *stemming*, uma vez que para a nuvem de palavras com os termos mais frequentes, são necessários os textos em seu formato completo, sem a redução das palavras aos seus radicais. Os resultados são apresentados a seguir. A figura 34 evidencia as palavras positivas associadas ao tweets da base de dados do Athlético Paranaense.

FIGURA 34 - NUVEM DE PALAVRAS ASSOCIADAS AOS TWEETS POSITIVOS DA BASE DE DADOS DO ATHLÉTICO PARANANESE


FONTE: O autor (2019).

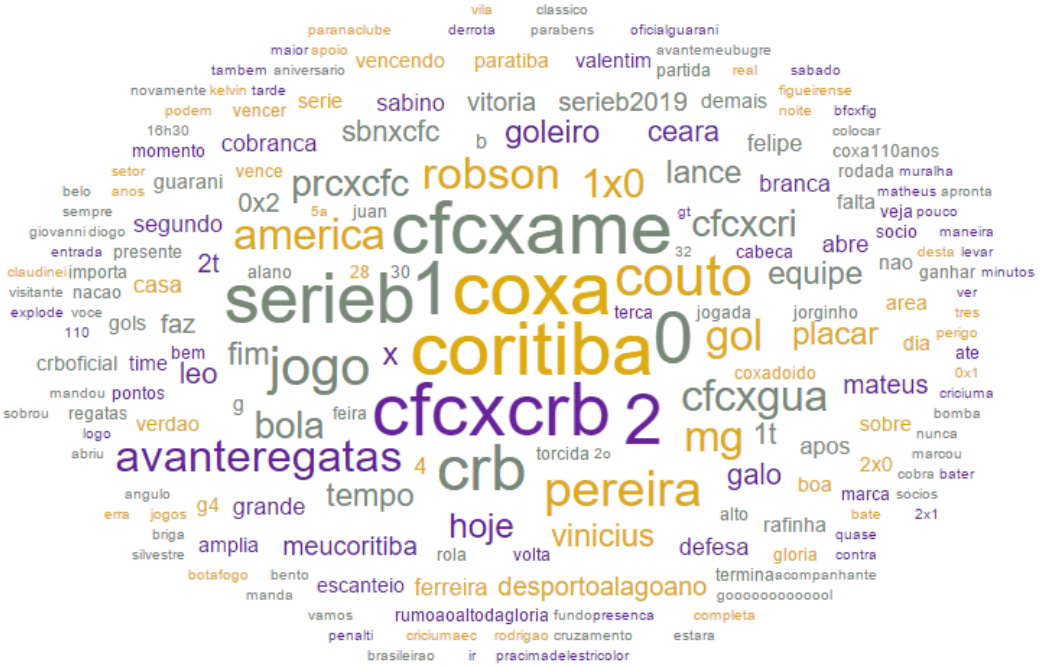
Como pode ser observado, palavras como “campeão”, “furacão campeão” e “viva o furacão” são palavras recorrentes na base de dados, devido ao título conquistado pelo clube durante a coleta dos dados desta pesquisa. Da mesma forma, nas palavras negativas também são recorrentes os termos, evidenciando a grande interação que houve no Twitter durante o período que o clube disputou o título da Copa do Brasil. Apesar disso, nas sentenças negativas é possível observar uma grande quantidade de termos da palavra “não”, aparecendo em destaque na figura 35.

FIGURA 35 - NUVEM DE PALAVRAS ASSOCIADAS AOS TWEETS NEGATIVOS DA BASE DE DADOS DO ATHLÉTICO PARANANESE


FONTE: O autor (2019).

Por sua vez, as palavras positivas do Coritiba com mais destaque são “Coritiba”, “Coxa” e “CFCxCRB”. Tweets como “Fazia tempo que não via meu Coxa jogar assim. Parece que o Jorginho realmente está conseguindo colocar ordem no time” ou “foi suado e sofrido, mas vencemos mais uma. Valeu meu #Verdão #CoxaBranca! ” foram determinantes para tais palavras se destacarem entre as mais populares, conforme demonstrado na figura 36.

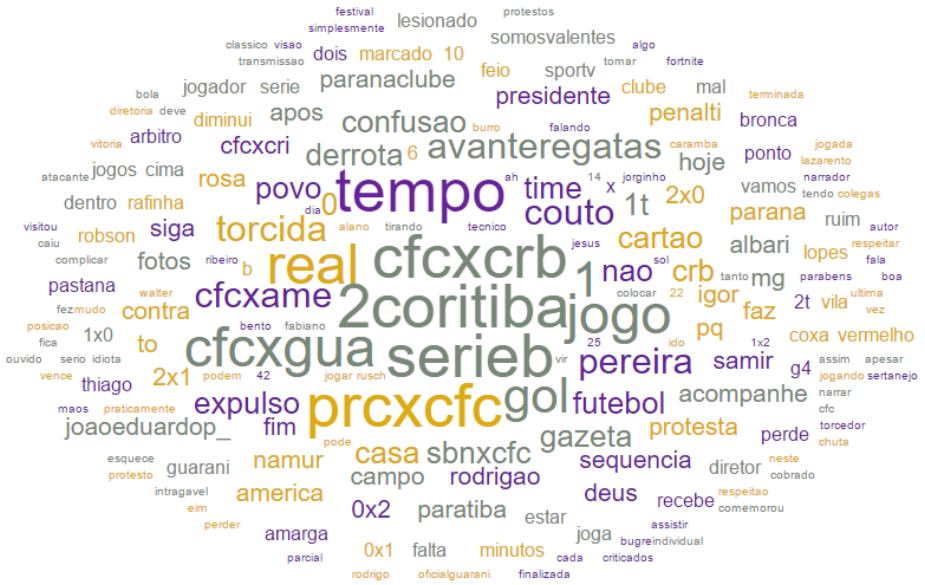
FIGURA 36 - NUVEM DE PALAVRAS ASSOCIADAS AOS TWEETS POSITIVOS DA BASE DE DADOS DO CORITIBA




FONTE: O autor (2019).

Palavras negativas com maior destaque na base de dados do Coritiba são “Coritiba”, “CFCxCRB”, “Série B”, “Tempo” e “PRCxCFC” e estão demonstradas na figura 37.

FIGURA 37 - NUVEM DE PALAVRAS ASSOCIADAS AOS TWEETS NEGATIVOS DA BASE DE DADOS DO CORITIBA



FONTE: O autor (2019).

Na base de dados do Paraná Clube, termos como “Paraná”, “Somos Valentes” e “gol” aparecem com maior destaque nos dados positivos, como em “#SérieB | #PRCxBRA | 1x0 | 3'/1T GOOOOOOL DO PARANÁ! Tá lá Tricolor! Com apenas três minutos abre o placar! ” ou ainda em “HOJE É DIA DE PARATIBA E NADA MAIS IMPORTA! Te espero na Vila. #PRCxCFC #SomosValentes #PRaCimaDelesTricolor” A Figura 38 demonstra a nuvem de palavras com os termos positivos mais populares retirados da base de dados do Paraná Clube.

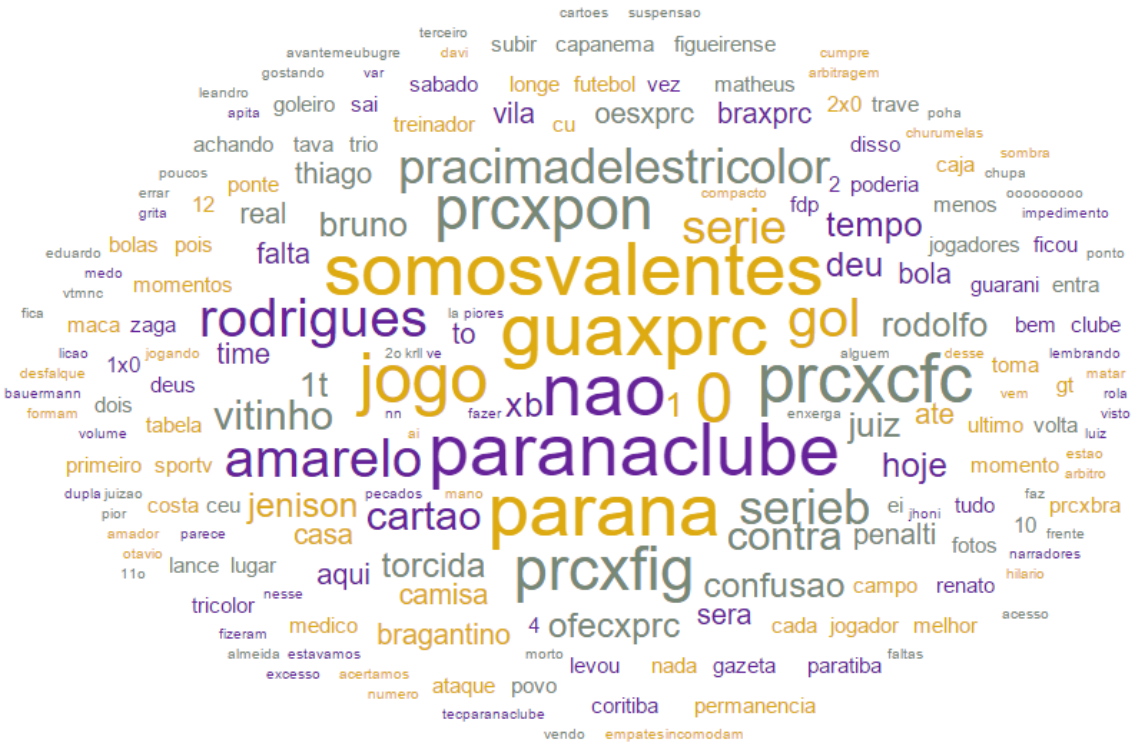
FIGURA 38 - NUVEM DE PALAVRAS ASSOCIADAS AOS TWEETS POSITIVOS DA BASE DE DADOS DO PARANÁ CLUBE


FONTE: O autor (2019).

Dentre os dados negativos, nota-se palavras como “não”, “jogo” e “amarelo” com bastante destaque, evidenciados por tweets como “ Rodolfo é desfalque na zaga hoje. Ele levou o terceiro amarelo e cumpre suspensão. Leandro Almeida e Eduardo Bauermann formam nossa dupla de zaga!” e “Alguém poderia fazer um compacto dos piores momentos do @ParanaClube no jogo #GUAxPRC. Me parece que nosso treinador @TecParanaClube só vê os poucos momentos que acertamos e não

enxerga nossos pecados” demonstram exemplos de sentenças negativas na base de dados. A figura 39 destaca as palavras associadas aos tweets negativos do Paraná Clube.

FIGURA 39 - NUVEM DE PALAVRAS ASSOCIADAS AOS TWEETS NEGATIVOS DA BASE DE DADOS DO PARANÁ CLUBE




FONTE: O autor (2019).

## 4.2 RESULTADOS DAS BASES ANALISADAS

Nesta seção estão evidenciados os experimentos realizados utilizando os algoritmos SVM, Naïve Bayes e CART. Nota-se a partir dos resultados uma alta taxa de acerto na classificação das instâncias, com um desempenho acima de 60% para todas as execuções realizadas. Conforme definido nos encaminhamentos metodológicos, a seleção de 500 registros aleatórios divididos na proporção de 70% dos dados para o treinamento dos algoritmos e 30% para a fase de teste dos modelos produziu uma baixa quantidade de dados utilizados pelos algoritmos, tornando rápida a execução dos algoritmos, retornando os resultados em sua maioria com um tempo inferior a 5 segundos. Os resultados demonstram que a análise de sentimentos pode

ser usada pelos clubes como uma ferramenta útil nas redes sociais, uma vez que os algoritmos classificam os dados com uma precisão considerável.

O algoritmo com o melhor desempenho geral foi o SVM, com uma média de acerto nas três bases de 71.66%. O Naive Bayes também teve uma alta taxa de acerto quando realizada uma média entre as três bases, totalizando um acerto de 71.33%. Apesar do algoritmo CART ter tido a menor taxa de instâncias classificadas corretamente dentre os três algoritmos utilizados, o desempenho foi considerado satisfatório, uma vez que classificou 68% dos dados corretamente. Nota-se uma taxa de acerto satisfatória para todos os algoritmos de aprendizado de máquina utilizados.

Ao verificar qual base de dados possuí a melhor média de acerto quando analisado todas as taxas de acerto, destaca-se a do Coritiba Football Clube, com 76.33% de acerto. As bases de dados do Athlético Paranaense e do Paraná Clube apresentaram 67.33% de acerto quando verificadas as médias dentre os três algoritmos. O quadro 10 evidencia as informações de acerto de cada base de dados, bem como a média de acerto entre as bases e os algoritmos.

QUADRO 10 - RELAÇÃO DA TAXA DE ACERTO DOS ALGOTIRMOS COM AS BASES DE DADOS

| | **TAXA DE ACERTO DOS ALGORITMOS** | | | |
|---|---|---|---|---|
| **CLUBE** | SVM | CART | NAIVE BAYES | MÉDIA |
| ATHLÉTICO | 70 | 66 | 67,33 | 67,77 |
| CORITIBA | 74 | 78 | 76,33 | 76,11 |
| PARANÁ | 71 | 60 | 67,33 | 66,11 |
| MÉDIA | 71,67 | 68 | 70,33 | |

FONTE: O autor (2019).

### 4.2.1 Resultados para a base de dados do Athlético Paranaense

A base de dados do Athlético Paranaense continha 2435 registros, dos quais selecionou-se 500 destes, de modo aleatório, para realizar a aplicação dos algoritmos de AM. Os resultados para os três algoritmos podem ser vistos na figura 40. O algoritmo Naive Bayes e CART obtiveram uma acurácia de 66% na classificação dos dados, enquanto o SVM retornou uma acurácia de 70% das instâncias classificadas corretamente, sendo o modelo que melhor classificou os dados.

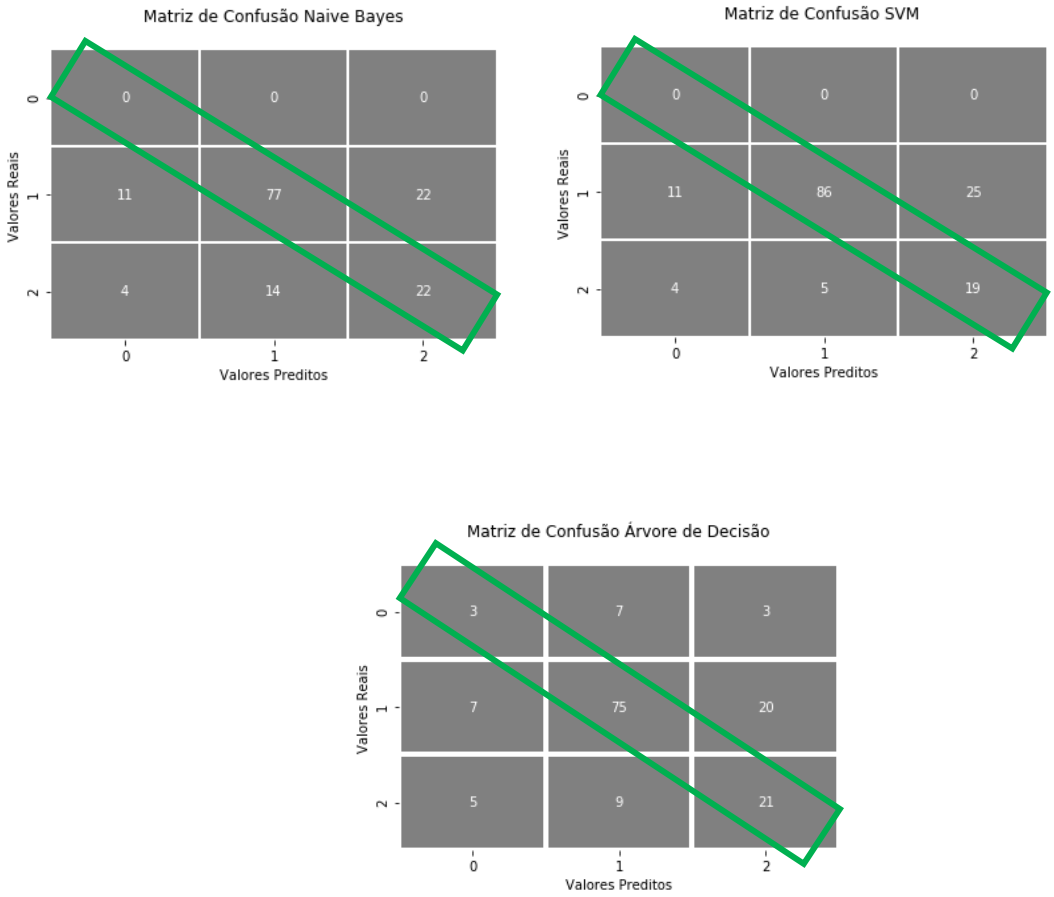
FIGURA 40 - RESULTADOS DOS ALGORITMOS DE CLASSIFICAÇÃO DE TEXTO APLICADOS NA BASE DE DADOS DO ATHLÉTICO PARANAENSE

```
Decision Tree Score =  66.0
              precision    recall  f1-score   support

    Positivo      0.200     0.231     0.214        13
      Neutro      0.824     0.735     0.777       102
    Negativo      0.477     0.600     0.532        35

    accuracy                          0.660       150
   macro avg      0.500     0.522     0.508       150
weighted avg      0.689     0.660     0.671       150

Naive Bayes Accuracy Score =  66.0
              precision    recall  f1-score   support

    Positivo      0.000     0.000     0.000         0
      Neutro      0.846     0.700     0.766       110
    Negativo      0.500     0.550     0.524        40

    accuracy                          0.660       150
   macro avg      0.449     0.417     0.430       150
weighted avg      0.754     0.660     0.702       150

SVM Accuracy Score =  70.0
              precision    recall  f1-score   support

    Positivo      0.000     0.000     0.000         0
      Neutro      0.945     0.705     0.808       122
    Negativo      0.432     0.679     0.528        28

    accuracy                          0.700       150
   macro avg      0.459     0.461     0.445       150
weighted avg      0.849     0.700     0.755       150
```

FONTE: O autor (2019).

Além da acurácia do modelo, a figura 40 também apresenta os valores de precisão, *recall* e *F1-Score* das instâncias classificadas. Outra forma de avaliação dos modelos é dada pela matriz de confusão de cada algoritmo, de modo que a partir da diagonal principal de cada matriz é possível verificar a quantidade de instâncias classificadas de acordo. Os resultados da matriz de confusão dos três algoritmos podem ser visualizados na figura 41. Nota-se nas matrizes de confusão dos algoritmos Naive Bayes e SVM a falta de resultados positivos. Contudo, para os valores do tipo “neutro”, foram classificadas corretamente 77 e 86 instâncias, respectivamente.

FIGURA 41 - MATRIZ DE CONFUSÃO DOS ALGORITMOS CLASSIFICADORES DOS DADOS DA BASE DO CLUBE ATHLÉTICO PARANAENSE




FONTE: O autor (2019).

### 4.2.2 Resultados para a base de dados do Coritiba Football Clube

Na base do Coritiba também foram selecionados 500 registros aleatórios, uma vez que a base possuía 545 instâncias rotuladas. Feita a seleção dos registros, utilizou-se novamente os algoritmos de aprendizado de máquina. A solução dos algoritmos está exibida na imagem 42.

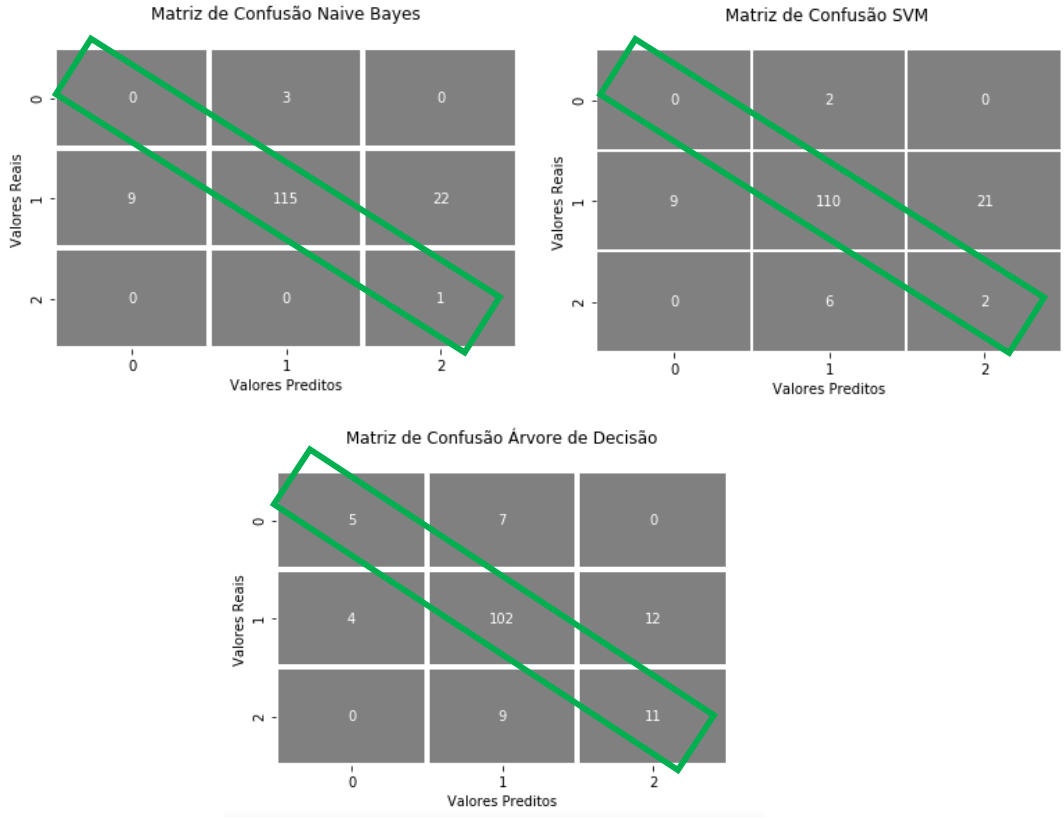
FIGURA 42 - RESULTADOS DOS ALGORITMOS DE CLASSIFICAÇÃO DE TEXTO APLICADOS NA BASE DE DADOS DO CORITIBA FOOTBALL CLUBE

```
Decision Tree Score =  78.66666666666666
              precision    recall  f1-score   support

    Positivo      0.556     0.417     0.476        12
      Neutro      0.864     0.864     0.864       118
    Negativo      0.478     0.550     0.512        20

    accuracy                          0.787       150
   macro avg      0.633     0.610     0.617       150
weighted avg      0.788     0.787     0.786       150

Naive Bayes Accuracy Score =  77.33333333333333
              precision    recall  f1-score   support

    Positivo      0.000     0.000     0.000         3
      Neutro      0.975     0.788     0.871       146
    Negativo      0.043     1.000     0.083         1

    accuracy                          0.773       150
   macro avg      0.339     0.596     0.318       150
weighted avg      0.949     0.773     0.849       150

SVM Accuracy Score =  74.66666666666667
              precision    recall  f1-score   support

    Positivo      0.000     0.000     0.000         2
      Neutro      0.932     0.786     0.853       140
    Negativo      0.087     0.250     0.129         8

    accuracy                          0.747       150
   macro avg      0.340     0.345     0.327       150
weighted avg      0.875     0.747     0.803       150
```

FONTE: O autor (2019).

Os resultados expõem que na base do Coritiba, ao contrário dos resultados obtidos na base do Athlético Paranaense, o algoritmo com o melhor desempenho foi o CART, com uma acurácia de 78.66%. Nota-se um alto acerto por parte dos três algoritmos utilizados, uma vez que os algoritmos Naive Bayes e SVM obtiveram uma acurácia de 77.33% e 74.66%, respectivamente. As métricas de precisão, *recall* e *F1-Score* evidenciaram resultados sólidos, principalmente na classificação de instâncias neutras, as quais compõem 74.68% da base de dados. Assim sendo, é possível notar um alto aprendizado por parte dos algoritmos, conforme demonstra as matrizes de confusão expostas na figura 43.

FIGURA 43 - MATRIZ DE CONFUSÃO DOS ALGORITMOS CLASSIFICADORES DOS DADOS DA BASE DO CORITIBA FOOTBALL CLUBE




FONTE: O autor (2019).

Nas três matrizes de confusão apresentadas, é possível notar uma quantidade de acerto superior a 100 instâncias, confirmando a alta assertividade dos algoritmos na base de dados do Coritiba.

### 4.2.3 Resultados para a base de dados do Paraná Clube

A base de dados do Paraná Clube conta com 435 registros e assim sendo, foi utilizada de maneira completa, selecionando todos os dados. Devida a uma quantidade menor de registros, houveram menos instâncias para treinar e testar os algoritmos. Apesar disso, nota-se que os resultados são semelhantes as demais bases de dados já apresentadas. A figura 44 evidencia os resultados obtidos com a aplicação dos algoritmos na base de dados do Paraná Clube.

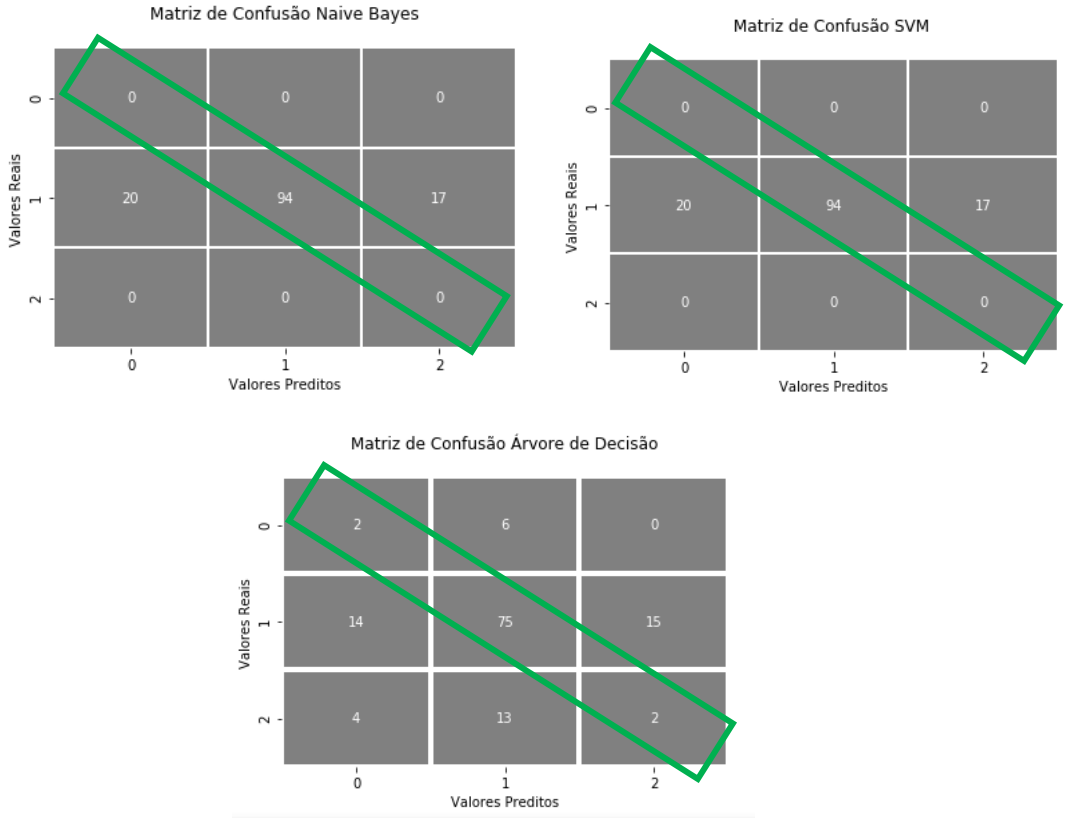
FIGURA 44 - RESULTADOS DOS ALGORITMOS DE CLASSIFICAÇÃO DE TEXTO APLICADOS NA BASE DE DADOS DO PARANÁ CLUBE

```
Decision Tree Score =  60.30534351145038
              precision    recall  f1-score   support

    Positivo      0.100     0.250     0.143         8
      Neutro      0.798     0.721     0.758       104
    Negativo      0.118     0.105     0.111        19

    accuracy                          0.603       131
   macro avg      0.339     0.359     0.337       131
weighted avg      0.657     0.603     0.626       131

Naive Bayes Accuracy Score =  71.7557251908397
              precision    recall  f1-score   support

    Positivo      0.000     0.000     0.000         0
      Neutro      1.000     0.718     0.836       131
    Negativo      0.000     0.000     0.000         0

    accuracy                          0.718       131
   macro avg      0.333     0.239     0.279       131
weighted avg      1.000     0.718     0.836       131

SVM Accuracy Score =  71.7557251908397
              precision    recall  f1-score   support

    Positivo      0.000     0.000     0.000         0
      Neutro      1.000     0.718     0.836       131
    Negativo      0.000     0.000     0.000         0

    accuracy                          0.718       131
   macro avg      0.333     0.239     0.279       131
weighted avg      1.000     0.718     0.836       131
```

FONTE: O autor (2019).

Ao observar a acurácia dos algoritmos, nota-se que a árvore de decisão (CART) apresentou o menor desempenho, com uma acurácia de 60%. Destaca-se ainda a taxa de acerto de 71.75% para os algoritmos Naive Bayes e SVM.

FIGURA 45 - MATRIZ DE CONFUSÃO DOS ALGORITMOS CLASSIFICADORES DOS DADOS DA BASE DO PARANÁ CLUBE




FONTE: O autor (2019).

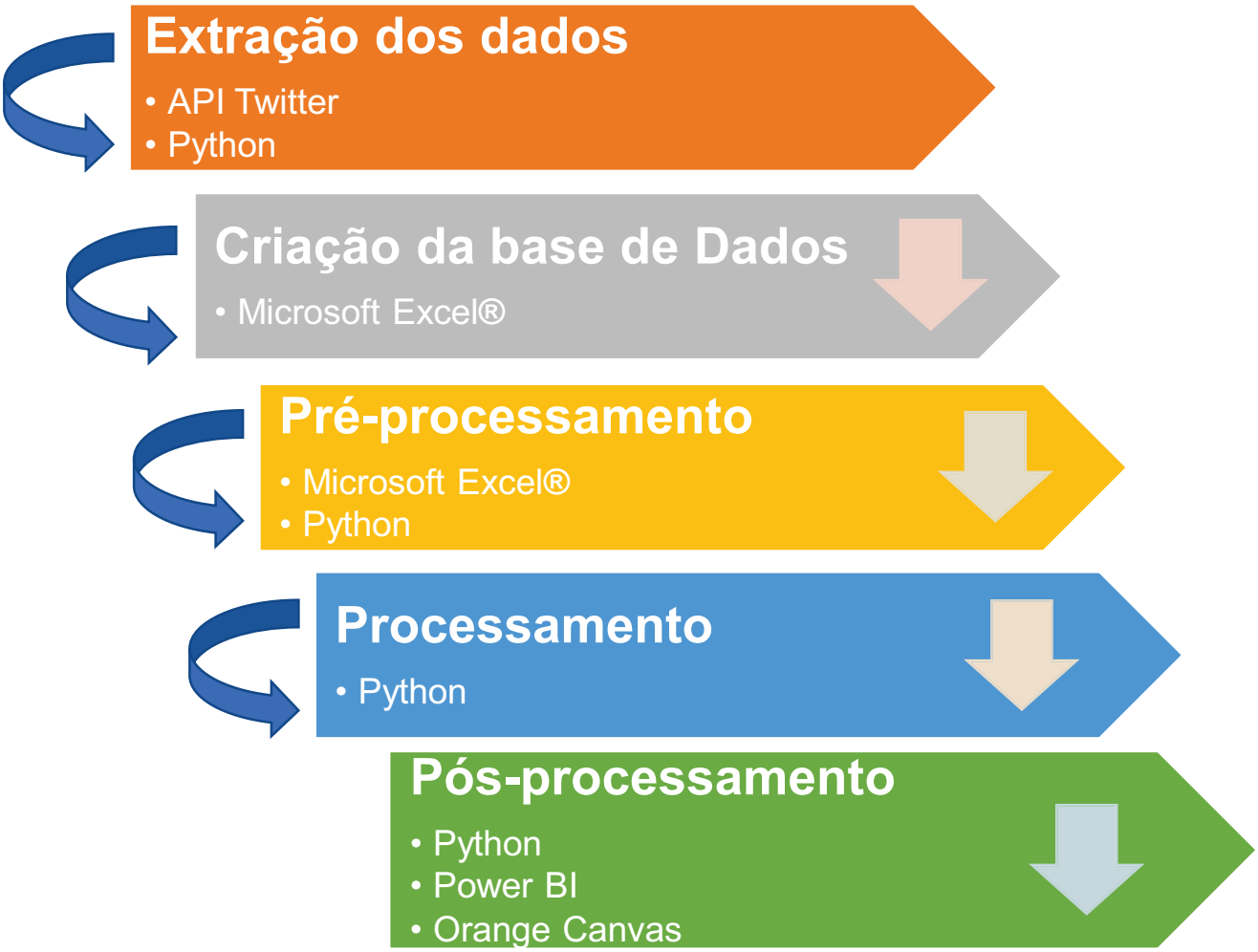
As matrizes de confusão retratadas na figura 45 confirmam uma grande disparidade entre instâncias neutras das demais, uma vez que instâncias negativas e positivas não foram utilizadas no conjunto de testes dos algoritmos Naive Bayes e SVM. Como destacado na Seção 4.1, a base de dados do Paraná Clube se destaca na porcentagem de registros neutros.

## 4.3 FLUXO DE ANÁLISE DE DADOS ESPORTIVOS NO TWITTER

Para atingir o objetivo específico de desenvolver um fluxo para coletar, tratar e processar os dados coletados do Twitter das equipes de futebol definidos nesta pesquisa, desenvolveu-se um fluxo abordando as etapas de extração dos dados, pré-processamento, processamento e pós-processamento da base, estruturando as principais etapas realizadas na metodologia desenvolvida nesta pesquisa para atingir os objetivos. A partir das Seções 3.2.1, 3.2.2, 3.2.3 e 3.2.4, criou-se um fluxo com

todas as etapas envolvidas na análise de sentimentos. A figura 46 exibe o processo e as ferramentas utilizadas durante cada uma das fases.

FIGURA 46 - FLUXOGRAMA DA METODOLOGIA UTILIZADA NA ANÁLISE DE SENTIMENTOS




FONTE: O autor (2019).

Este fluxo serve principalmente no contexto da análise de dados esportivas nas redes sociais, pois detalha as etapas realizadas para a realização da análise de sentimentos de clubes de futebol, uma vez que apresenta as ferramentas utilizadas, as bibliotecas implementadas pela linguagem de programação Python e a utilização da API do Twitter para extrair dados dos clubes observados nesta pesquisa.

Na etapa de extração dos dados, a API foi utilizada para realizar a conexão entre o código desenvolvido em linguagem de programação Python com a rede social Twitter e juntamente com um código em linguagem Python, extrair os dados.

Feita a extração, os dados foram estruturados e agrupados em uma planilha, facilitando a organização dos dados. A planilha serviu como fonte de dados e também

para etapas de pré-processamento, que foi feita majoritariamente utilizando a linguagem *Python*.

Com os dados normalizados, a etapa de processamento ocorreu também em linguagem Python, empregando os algoritmos de aprendizado de máquina para classificar os sentimentos dos torcedores.

Para apresentar os dados na etapa de pós-processamento, utilizou-se ainda a linguagem Python com o intuito de apresentar as saídas dos algoritmos de aprendizado de máquina, além dos softwares Orange Canvas na criação da nuvem de palavras e Power BI para a exposição dos dados em formatos gráficos.

O fluxo apresentado nesta etapa, juntamente com a explanação a respeito das ferramentas e bibliotecas utilizadas serve para auxiliar novos trabalhos em análise de sentimentos em clubes de futebol, pois detalha as ações necessárias em cada etapa, como a criação da API no *Twitter*, a definição da recuperação dos dados por meio das hashtags criadas pelos clubes, a utilização de bibliotecas em linguagem de programação *Python* e a criação de nuvens de palavras. Vale ressaltar que outras áreas podem se beneficiar do fluxo apresentado, mas com adaptações para a realidade estudada.

Feita a apresentação do fluxograma bem como das ações realizadas em cada etapa, são apresentadas as considerações finais desta pesquisa, salientando a validação dos objetivos propostos, as principais contribuições, limitações da pesquisa e trabalhos futuros.

# 5 CONSIDERAÇÕES FINAIS

Clubes de futebol estão organizados como instituições complexas que necessitam tomar decisões inteligentes a partir de informações confiáveis. Os avanços tecnológicos, a popularização das redes sociais e, consequentemente, as expansões no volume de dados permitem aos clubes novas formas de obter e gerenciar informações, tornando-as úteis no processo decisório.

A mineração de textos auxilia na análise dos dados nas redes sociais, o que pode beneficiar instituições esportivas que buscam maior aproximação com seus torcedores. A análise de sentimentos nas redes sociais alicerça a execução de estratégias e posicionamentos de clubes de futebol, uma vez que é possível verificar a satisfação de seus torcedores e associados.

Optou pela análise de sentimentos em clubes de futebol pelo fato de serem instituições extremamente populares e responsáveis por movimentarem postagens, debates e assuntos nas redes sociais. A partir das coletas dos dados, ficou evidente como instituições esportivas despertam interesse nas redes sociais e geram uma alta quantidade de postagens.

O maior desafio desta pesquisa residiu nas etapas de pré-processamento, uma vez que os dados coletados do Twitter possuem peculiaridades como falta de preocupação com pontuação e escrita, ironias, *emoticons,* imagens, vídeos e compartilhamentos de postagens de outros usuários. Tais características dificultaram um pré-processamento com melhor desempenho e, consequentemente, dificultaram uma análise mais assertiva dos algoritmos de aprendizado de máquina. A entrevista também apresentou desafios, uma vez que contatar as instituições estudadas nesta pesquisa provou-se um processo demorado e sem sucesso em grande parte dos veículos de comunicação utilizados. Contudo, a entrevista realizada com o Clube Athlético Paranaense corroborou com a afirmação de que instituições esportivas como clubes de futebol podem ser beneficiar da utilização de técnicas de análise de sentimentos.

A partir da coleta dos dados e das aplicações de aprendizado de máquina, ficou evidente como a mineração de opiniões pode auxiliar clubes de futebol em analisar postagens de seus torcedores na rede social.

## 5.1 VERIFICAÇÃO DOS OBJETIVOS PROPOSTOS COM OS RESULTADOS

O objetivo geral desta pesquisa baseou-se em **identificar e analisar os sentimentos dos torcedores dos três principais clubes de futebol do Estado do Paraná em suas publicações no Twitter.** Para alcançar este objetivo, definiu-se três objetivos específicos: a) coletar dados no Twitter referentes aos clubes selecionados; b) desenvolver uma metodologia para coletar, tratar e processar os dados coletados do Twitter; c) aplicar algoritmos de aprendizado de máquina, buscando classificar os tweets de acordo com o sentimento contido nele.

O primeiro objetivo específico, **coletar dados no Twitter referentes aos clubes selecionados**, foi atingido durante a etapa de montagem da base de dados, de modo que após o final de cada jogo, os tweets referentes as partidas foram coletados e organizados em uma base de dados. Uma quantidade maior de partidas necessárias para a criação da base de dados dos clubes Coritiba e Paraná comprovam uma interação nas redes sociais pouco expressivas, dificultando a extração dos dados e a montagem das bases de dados.

O segundo objetivo específico, **desenvolver um fluxo para coletar, tratar e processar os dados coletados do Twitter,** foi atingido por meio das Seções 3.2.1, 3.2.2, 3.2.3 e 3.2.4 e sumarizado na seção 4.3. Nas Seções descritas, buscou-se criar um método estruturado para coletar e realizar as etapas de pré-processamento, processamento e pós-processamento dos dados. Por meio do fluxo, é possível realizar todas as etapas de tratamento dos dados para aplicação nos algoritmos de aprendizado de máquina. O método inicialmente buscou criar uma base de dados com as postagens feitas pelos torcedores dos clubes analisados e, após a montagem da base, aplicar técnicas de pré-processamento a fim de facilitar a aplicação dos algoritmos de aprendizado de máquina responsáveis por classificar os dados. Com a aplicação dos algoritmos realizadas, a próxima etapa do fluxo foi a análise dos resultados, evidenciando a assertividade de cada modelo de aprendizado de máquinas nas três bases desenvolvidas para esta pesquisa.

O terceiro objetivo específico, **aplicar algoritmos de aprendizado de máquina, buscando classificar os tweets de acordo com o sentimento contido nele**, foi alcançado por meio da execução dos algoritmos de aprendizado supervisionado Naive Bayes, SVM e Árvore de Decisão (CART), conforme apresentados na Seção 4. A aplicação dos algoritmos permitiu verificar a quantidade

de instâncias classificadas corretamente, bem como verificar a capacidade dos algoritmos em definir o sentimento contido dentro dos tweets. A classificação correta de 71.66% das instâncias com o algoritmo SVM, 71.33% com o algoritmo Naive Bayes e 68% com o algoritmo CART evidenciam como é possível utilizar técnicas de aprendizado de máquina para classificar os sentimentos das publicações realizadas pelos torcedores no Twitter.

Desta forma, por meio dos resultados alcançados pelos cumprimentos dos objetivos específicos, julga-se que a pesquisa atingiu o objetivo geral proposto, que consistia em identificar e analisar os sentimentos dos torcedores dos três principais clubes de futebol do Estado do Paraná em suas publicações no Twitter.

## 5.2 CONTRIBUIÇÕES DO TRABALHO

A contribuição mais notável deste trabalho consiste na apresentação de um fluxo para coletar, tratar e avaliar dados referentes aos tweets de torcedores de clubes de futebol. O fluxo criado durante os encaminhamentos metodológicos pode ser utilizado por outras pesquisas na área, uma vez que o fluxo demonstra as ferramentas utilizadas em cada etapa. Com a adaptação de um fluxo ao contexto da análise de sentimentos, espera-se facilitar novas pesquisa na área, além de exibir a aplicação das técnicas de análise de sentimentos e recuperação da informação no contexto esportivo. A demonstração das bibliotecas utilizadas na linguagem de programação Python juntamente com as demais aplicações facilita outros trabalhos com análise de sentimentos. Espera-se que este trabalho possa contribuir para a área de *marketing* e de relações externas dos clubes, auxiliando no entendimento de como é possível analisar os sentimentos dos torcedores e verificar os assuntos discutidos pelos torcedores dentro das redes sociais. A aplicação desta pesquisa pode ser realizada em diversos clubes de futebol e por um período de tempo maior, possibilitando uma análise mais profunda e robusta. Ainda é possível aplicar outras formas de avaliações, buscando descobrir a influência de cada usuário que realizou uma postagem, a localização dos tweets, a data de cada postagem, além da quantidade de tweets que foram replicados na forma de *retweets*.

Outra contribuição deste trabalho foi a verificação dos sentimentos dos torcedores em tweets de língua portuguesa, uma vez que a literatura existente neste meio ainda é limitada. Espera-se que as etapas desenvolvidas nesta pesquisa possam

servir para outros pesquisadores que necessitem classificar e analisar sentimentos no idioma português, especialmente em instituições esportivas.

## 5.3 LIMITAÇÕES DA PESQUISA E TRABALHO FUTUROS

Esta pesquisa apresenta limitações, resultantes principalmente da rotulação dos dados, uma vez que rotular dados manualmente é um processo complexo e que consome muito tempo (LIU, LI e GUO, 2012, p. 1678). Além disso, a rotulação foi realizada somente por uma pessoa, fato que pode afetar a classificação dos dados. Sugere-se na etapa de rotulação diferentes abordagens e a utilização de aplicativos capazes de avaliar sentimentos de cada sentença ou até mesmo a classificação manual dos dados realizada por um número maior de pessoas, a fim de evitar erros na rotulação dos tweets.

Outra limitação consistiu na avaliação dos sentimentos de somente três clubes de futebol, fator limitador para análises mais complexas e, consequentemente, ideias mais profundas. Sugere-se, desta forma, coletar e avaliar sentimentos de um número maior de times de futebol, buscando estabelecer novas relações entre os dados. Também é possível desenvolver pesquisas futuras coletando dados de um campeonato específico ou grandes eventos específicos, como a Copa do Mundo FIFA® ou a UEFA *Champions League*®, eventos mundialmente conhecidos e que atraem milhares de espectadores.

Adicionalmente, sugere-se aplicar a análise de sentimentos nos clubes de futebol em outras redes sociais, como Facebook, Instagram e YouTube. Podem ser realizadas comparações entre as redes sociais, estudando o comportamento dos torcedores em diferentes redes sociais e verificando características em cada uma delas.

# REFERÊNCIAS


ALAMAR, Benjamin; MEHROTRA, Vijay. Beyond moneyball: The rapidly evolving world of sports analytics. **Analytics Magazine**, Catonsville, p.33-37, out. 2011. Disponível em: <http://analytics-magazine.org/beyond-moneyball-the-rapidly-evolving-world-of-sports-analytics-part-i-2/>. Acesso em: 21 mar. 2019.

ALVES, André et al. a comparison of svm versus naive-bayes techniques for sentiment analysis in tweets. **Proceedings Of The 20th Brazilian Symposium On Multimedia And The Web - Webmedia '14**, [s.l.], p.123-130, 2014. ACM Press.

AMIDI, Afshine; AMIDI, Shervine. **Machine learning tips and tricks cheatsheet.** Disponível em: <https://stanford.edu/~shervine/teaching/cs-229/cheatsheet-machine-learning-tips-and-tricks>. Acesso em: 16 out. 2019.

ARANHA, Christian; PASSOS, Emmanuel. A Tecnologia de Mineração de Textos. **Revista Eletrônica de Sistemas de Informação**, [s.l.], v. 5, n. 2, p.1-8, 31 ago. 2006. IBEPES (Instituto Brasileiro de Estudos e Pesquisas Sociais).

ARMAH, Gabriel Kofi; LUO, Guangchun; QIN, Ke. A Deep Analysis of the Precision Formula for Imbalanced Class Distribution. **International Journal of Machine Learning and Computing**, [s.l.], v. 4, n. 5, p.417-422, 2014. EJournal Publishing.

ASSIS, Wilson Martins de. **Gestão da informação nas organizações.** SI: Autêntica Editora, 2008.

BAEZA-YATES, Ricardo; RIBEIRO-NETO, Berthier. **Modern information retrieval.** New York: Acm Press, 1999.

BARION, Eliana Cristina Nogueira; LAGO, Decio. Mineração de textos. **Revista de Ciências Exatas e Tecnologia**, Valinhos, v. 3, n. 3, p.123-140, dez. 2008.

BARNAGHI, Peiman; GHAFFARI, Parsa; BRESLIN, John G. Opinion mining and sentiment polarity on twitter and correlation between events and sentiment. **2016 ieee second international conference on big data computing service and applications (bigdataservice)**, [s.l.], p.52-57, mar. 2016. IEEE.

BOYD, Danah; GOLDER, Scott; LOTAN, Gilad. Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter. In: HAWAII INTERNATIONAL CONFERENCE ON SYSTEM SCIENCES, 43., 2010, Kauai. **Proceedings**. IEEE, 2010. p. 1 - 10. Disponível em: <https://ieeexplore.ieee.org/document/5428313>. Acesso em: 01 set. 2019.

BREIMAN, Leo et al.. **Classification and regression trees.** Belmont: Wadsworth International Group, 1984.

CASTRO, Leandro Nunes de; FERRARI, Daniel Gomes. **Introdução à mineração de dados.** São Paulo: Saraiva, 2016.

CLARK, Alexander; FOX, Chris; LAPPIN, Shalom. **The handbook of computational linguistics and natural language processing**. Oxford: Wiley-blackwell, 2010. Disponível em: <http://course.duruofei.com/wp-content/uploads/2015/05/Clark_Computational-Linguistics-and-Natrual-Language-Processing.pdf>. Acesso em: 04 out. 2019

CHOO, Chun Wei. **A organização do conhecimento**: como as organizações usam a informação para criar significado, construir conhecimento e tomar decisões. Tradução Eliana Rocha. - São Paulo: Editora Senac São Paulo, 2003.

CUNNINGHAM, S.J.; LITTIN, J.N.; WITTEN, I.H. **Applications of machine learning in information retrieval.** Hamilton: University Of Waikato, 1997. Disponível em: <https://www.cs.waikato.ac.nz/~ml/publications1997_bib.html#Cunningham1997_4>. Acesso em: 12 jun. 2019.

DAVENPORT, Thomas H. **Ecologia da informação**: por que só a tecnologia não basta para o sucesso na era da informação. São Paulo: Futura, 1998. 312 p. (ISBN 85-86082-72-4) Tradução Bernadette Siqueira Abrão.

FELDMAN, Ronen; SANGER, James. **The text mining handbook:** Advanced Approach in Analyzing Unstructured Data. New York: Cambridge University Press, 2007.

FIGUEIREDO, Diego. **A profissionalização das organizações do futebol**: um estudo de casos sobre estratégia, estrutura e ambiente dos clubes brasileiros**.** 2011. 264 f. Dissertação (Mestrado) - Curso de Administração, Centro de Pós-graduação e Pesquisas em Administração, Universidade Federal de Minas Gerais, Belo Horizonte, 2011.

FILHO, José Adail Carvalho. **Mineração de textos**: análise de sentimento utilizando tweets referentes à Copa do Mundo 2014. 2014. 44 f. TCC (Graduação) - Curso de Engenharia de Software, Universidade Federal do Ceará, Quixadá, 2014. Disponível em: <http://www.repositoriobib.ufc.br/000017/0000179f.pdf>. Acesso em: 18 maio 2019.

FOIT, Antonio José Hable. **Descrição do processo de análise de opiniões no twitter:** bitcoin no cenário brasileiro. 2017. 72 f. TCC (Graduação) - Curso de Gestão da Informação, Universidade Federal do Paraná, Curitiba, 2017.

GIL, Antonio Carlos. **Como elaborar projetos de pesquisa.** 4. ed. São Paulo: Atlas S.a, 2002.

GONÇALVES, Emerson. **Os tamanhos e os números gigantes do futebol.** 2016. Disponível em: <http://globoesporte.globo.com/blogs/especial-blog/olhar-cronico-esportivo/post/os-tamanhos-e-os-numeros-gigantes-do-futebol.html>. Acesso em: 23 maio 2019.

IGLESIAS, Luis Sancliment. **A mineração de opinião em mídias sociais como ferramenta para medir a (in) satisfação do consumidor.** 2018. 124 f. TCC (Graduação) - Curso de Gestão da Informação, Universidade Federal do Paraná,

Curitiba, 2018. Disponível em: <https://acervodigital.ufpr.br/bitstream/handle/1884/59523/Luis%20Sancliment%20Iglesias.pdf?sequence=1&isAllowed=y>. Acesso em: 05 jun. 2019.

JIVANI, Anjali G. A Comparative Study of Stemming Algorithms**. Int. J. Comp. Tech. Appl**, Gujarat, v. 2, n. 1, p.1930-1938, dez. 2011. Disponível em: <https://pdfs.semanticscholar.org/1c0c/0fa35d4ff8a2f925eb955e48d655494bd167.pdf>. Acesso em: 04 out. 2019.

JUPYTER. **About us**. Disponível em: <https://jupyter.org/about>. Acesso em: 10 out. 2019.

KELLEHER, John D.; NAMEE, Brian Mac; D'ARCY, Aoife. **Fundamentals of machine learning for predictive data analytics:** algorithms, worked examples, and case studies**.** Cambridge: Mit Press, 2015.

KING, John; MAGOULAS, Roger. **2013 Data science salary survey:** Tools, Trends, What Pays (and What Doesn't) for Data Professionals. Sebastopol: O'reilly Media, 2014.

LEXALYTICS. **Semantria for Excel.** Disponível em: <https://www.lexalytics.com/semantria/excel>. Acesso em: 15 nov. 2019.

LIN, Jimmy; KOLCZ, Alek. Large-scale machine learning at twitter. **Proceedings of The 2012 International Conference On Management of Data**, p.793-804, 2012. ACM Press.

LIU, Bing. **Sentiment analysis and opinion mining**. Chicago: Morgan & Claypool Publishers, 2012.

LIU, K.; LI, W.; GUO, M. Emoticon Smoothed Language Models for Twitter Sentiment Analysis. **AAAI Conference on Artificial Intelligence**, North America, jul. 2012. Disponível em: <https://www.aaai.org/ocs/index.php/AAAI/AAAI12/paper/view/5083/5319>.

LUCAS, G et al. GOAALLL!: Using sentiment in the World Cup to explore theories of emotion. **2015 International Conference On Affective Computing and Intelligent Interaction (acii)**, [s.l.], p.898-903, set. 2015. IEEE.

MANRAI, Ajay K Et al.. Social Media: Past, Present and Future. **Routledge Companion on the Future of Marketing**. p. 234-249. 2013

MARCACINI, Ricardo M.; MOURA, Maria F.; REZENDE, Solange O.; O uso da Mineração de Textos para Extração e Organização Não Supervisionada de Conhecimento. **Revista de Sistemas de Informação da FSMA**, Macaé, v. 7, n. 1, p.7-21, jul. 2011.

MATTAR, Michel Fauze; MATTAR, Fauze Najib (Org.). **Gestão de negócios esportivos.** Rio de Janeiro: Elselvier, 2013.

MCCALLUM, Andrew; NIGAN, Kamal. A Comparison of Event Models for Naive Bayes Text Classification. **Work Learn Text Categ**, Pittsburgh, v. 752, n. 1, maio 2001.

MCKINNEY, Wes. **Python for data analysis:** Data Wrangling with Pandas, Numpy and IPython. 2. ed. Sebastopol: O'reilly Media, 2018.

MICROSOFT. **Microsoft Excel.** Disponível em: <https://products.office.com/en/excel>. Acesso em: 15 nov. 2019.

MICROSOFT. **O que é o Power BI?** Disponível em: <https://powerbi.microsoft.com/pt-pt/what-is-power-bi/>. Acesso em: 05 nov. 2019.

MLADENIC, D. e GROBELNIK, M. 1998. Word sequences as features in text-learning. In **Proceedings** of ERK-98, the Seventh Electrotechnical and Computer Science Conference, pp. 145–148, Ljubljana, SL. Disponível em: <https://pdfs.semanticscholar.org/8317/e01a4b7c94d5138d14ab3a4cf77a42977e89.pdf>. Acesso em: 10 abr. 2019.

MURPHY, Kevin P. **Machine learning**: a probabilistic perspective. Cambridge: Massachusetts Institute of Technology, 2012.

NLTK. **Examples for portuguese processing.** Disponível em: <http://www.nltk.org/howto/portuguese_en.html>. Acesso em: 10 out. 2019.

Osisanwo, F.Y et al.. Supervised Machine Learning Algorithms: Classification and Comparison. **International Journal of Computer Trends and Technology**, [s.l.], v. 48, n. 3, p.128-138, 25 jun. 2017. Seventh Sense Research Group Journals.

PANG, Bo; LEE, Lillian. Opinion mining and sentiment analysis. **Foundations and Trends® in Information Retrieval**, [s.l.], v. 2, n. 12, p.1-135, 2008. Now Publishers.

PATEL, Vishakha; PRABHU, Gayatri; BHOWMICK, Kiran. A Survey of Opinion Mining and Sentiment Analysis. **International Journal of Computer Applications**, [s.l.], v. 131, n. 1, p.24-27, 17 dez. 2015. Foundation of Computer Science.

Pedregosa, F et al. Scikit-learn: Machine Learning in {P}ython. **Journal of Machine Learning Research**, Si, v. 212, n. 1, p.2825-2830, 2011.

PLURI PESQUISAS ESPORTIVAS. **As maiores torcidas de futebol do Brasil em 2012.** Disponível em: <https://www.campeoesdofutebol.com.br/maiores_torcidas_pluri_2012.html>. Acesso em: 13 jun. 2019.

PRODANOV, Cleber Cristiano; FREITAS, Ernani Cesar de. **Metodologia do trabalho científico**: métodos e técnicas da pesquisa e do trabalho acadêmico**.** 2. ed. Novo Hamburgo: Feevale, 2013.

REVISTA FORBES. **Os 50 times de futebol mais valiosos da América em 2018.** Disponível em: <https://forbes.uol.com.br/listas/2018/08/os-50-times-de-futebol-mais-valiosos-da-america-em-2018/#foto11>. Acesso em: 23 maio 2019.

ROBERTSON, Stephen. Understanding Inverse Document Frequency: On theoretical arguments for IDF. **Journal of Documentation - J Doc**, London, v. 60, n. 1, p.503-520, out. 2004. Disponível em: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.438.2284&rep=rep1&type=pdf>. Acesso em: 12 out. 2019.

RODRIGUES, Alan Cristian Falcoski. **Modelo para análise de sentimentos no facebook:** um estudo de caso na página do Senado Federal Brasileiro. 2017. 83 f. TCC (Graduação) - Curso de Gestão da Informação, Universidade Federal do Paraná, Curitiba, 2017.

RUDNICK, Fernando. **Pesquisa revela que torcida do Atlético ampliou vantagem sobre a do Coritiba.** 2017. Disponível em: <https://www.gazetadopovo.com.br/esportes/futebol/pesquisa-revela-que-torcida-do-atletico-ampliou-vantagem-sobre-a-do-coritiba-8l0fgdbewiztjj86czyzw4rwb/>. Acesso em: 03 jun. 2019.

RUSSEL, Mathew A. **Mining the social web**: data mining Facebook, Twitter, LinkedIn, Google+, GitHub and more. 2 ed. Sebastopol: O'reilly Media, Inc., 2013.

SAIF, Hassan; FERNANDEZ, Miriam; ALANI, Harith. Automatic Stopword Generation using Contextual Semantics for Sentiment Analysis of Twitter. In: INTERNATIONAL SEMANTIC WEB CONFERENCE, 13., 2014, Riva del Garda. **Proceedings.** ISWC 2014, 2014. p. 281 - 284.

SATHYA, R.; ABRAHAM, Annamma. Comparison of supervised and unsupervised learning algorithms for pattern classification. **International Journal Of Advanced Research In Artificial Intelligence.** Bangalore, p. 34-38. jan. 2013. Disponível em: <http://ijarai.thesai.org/Downloads/IJARAI/Volume2No2/Paper_6-Comparison_of_Supervised_and_Unsupervised_Learning_Algorithms_for_Pattern_Classification.pdf>. Acesso em: 11 abr. 2019

SHALEV-SHWARTZ, Shai; BEN-DAVID, Shai. **Understanding machine learning from theory to algorithms.** New York: Cambridge University Press, 2014.

SILVA, Rajitha Minusha. **Sports analytics.** 2016. 64 f. Tese (Doutorado) - Curso de Filosofia, Rajarata University Of Sri Lanka, Mihintale, 2016. Disponível em: <https://www.stat.sfu.ca/content/dam/sfu/stat/alumnitheses/2016/Silva%2C%20Rajitha.pdf>. Acesso em: 22 mar. 2019.

SMAAL, Batriz. **A história do Twitter.** 2010. Disponível em: <https://www.tecmundo.com.br/rede-social/3667-a-historia-do-twitter.htm>. Acesso em: 16 jun. 2019.

TALIB, Ramzan et al.. Text Mining: techniques, applications and issues. **International Journal of Advanced Computer Science and Applications.** Faisalabad, p. 414-418. nov. 2016.

Tan, A., Mui, H.W., & Terrace, K. **Text mining:** the state of the art and the challenges**.** 2000. Disponível em: <https://www.semanticscholar.org/paper/Text-Mining%3A-The-state-of-the-art-and-the-Tan-Mui/52931e51dc9ea7317a8157ec49da7fb36cb364a4>. Acesso em: 20 maio 2019.

TORRES, Cláudio. **A bíblia do marketing digital**. São Paulo: Editora Novatec, 2009.

Twitter. **About Twitter's APIs.** Disponível em: <https://help.twitter.com/en/rules-and-policies/twitter-api>. Acesso em: 24 nov. 2019.

TWITTER. **The 2014 #YearOnTwitter**. 2014. Disponível em: <https://blog.twitter.com/en_us/a/2014/the-2014-yearontwitter.html>. Acesso em: 15 maio 2019.

VALENTINI, Giorgio; DIETTERICH, Thomas G. Bias-variance analysis of support vector machines for the development of SVM-based ensemble methods. **Journal of Machine Learning Research**, New York, v. 5, n. 1, p.725-775, jul. 2004. Disponível em: <http://www.jmlr.org/papers/v5/valentini04a.html>. Acesso em: 10 out. 2019.

VANDERPLAS, Jake. **Python data science handbook.** Sebastopol: O'reilly, 2017.

WEKA. **Weka 3:** machine learning software in Java**.** 2019. Disponível em: <https://www.cs.waikato.ac.nz/ml/weka/>. Acesso em: 04 jun. 2019.

ZANINI, Nadir; DHAWAN, Vikas. Text Mining: An introduction to theory and some applications. **Research Matters**, Si, v. , n. , p.38-44, jan. 2015.

## APÊNDICE 1 – ROTEIRO DE ENTREVISTA APLICADO NO CLUBE ATHLÉTICO PARANAENSE

(continua)

Universidade Federal do Paraná
Curso: Gestão da Informação
Trabalho de Conclusão de Curso
Orientadora: Denise Tsunoda
Aluno: Gabriel Luis Degani

---

### ROTEIRO DE ENTREVISTA

Pedimos sua participação nesta breve entrevista, a qual serve como um objeto de estudo para o trabalho de conclusão de curso do aluno Gabriel Degani, o qual visa estudar a mineração de opiniões nas redes sociais dentro dos clubes de futebol. As informações disponibilizadas nesta entrevista serão usadas somente para fins de estudo, não sendo utilizadas para obter nenhum tipo de vantagem comercial, financeira ou social. A duração desta entrevista é de, no máximo, trinta minutos, sendo um instrumento rápido de verificação.
A entrevista proposta tem como objetivo verificar questões relacionadas com a mineração de opiniões nas redes sociais dos clubes de futebol, de modo a investigar se existe um departamento relacionado com a área e de que forma as redes sociais podem impactar nas tomadas de decisões.

1. **Existe um departamento dentro do clube responsável pela recuperação da informação nas redes sociais e sua análise? Se a resposta for sim:**
    a. **Como é realizada tal análise?**
    b. **Qual(is) rede(s) social(is) é/são analisada(s)?**
    c. **Quantas pessoas integram o departamento?**
    d. **Qual a formação dessas pessoas?**
    e. **Há quanto tempo este trabalho é desenvolvido?**

    **Caso não exista um departamento específico dentro do clube, existe alguma empresa terceirizada responsável por essa gestão? Ainda que não exista, o clube utiliza as redes sociais para alguma finalidade? Quais?**

(conclusão)

2. **Você acredita que de algum modo as redes sociais podem contribuir para tomar decisões? Se sim, de qual forma?**

3. **As redes sociais influenciam em algum aspecto nas ações do clube? Quais ações já foram tomadas utilizando as redes sociais?**

## APÊNDICE 2 – TERMO DE PARTICIPAÇÃO DA ENTREVISTA

Universidade Federal do Paraná
Curso: Gestão da Informação
Trabalho de Conclusão de Curso
Orientadora: Denise Tsunoda
Aluno: Gabriel Luis Degani

---

### TERMO DE PARTICIPAÇÃO

A sua participação neste projeto de pesquisa é estritamente voluntária. Além disso, sr.(a) pode parar de responder a qualquer momento, sem precisar justificar.

Benefícios: Ao participar desta pesquisa ao sr. (a) não terá nenhum benefício direto. Não haverá nenhum valor econômico, a receber ou a pagar, pela participação nesta pesquisa. Se o sr.(a) sentir algum desconforto ou constrangimento para responder alguma pergunta, poderá recusar a responder a pesquisa.

Confidencialidade: A sua participação no referido estudo será no sentido de responder de forma espontânea às perguntas do questionário. O pesquisador compromete-se a proteger as informações pessoais obtidas e garantir a segurança dos dados dos participantes. A sua privacidade será respeitada, ou seja, o seu nome ou qualquer outro dado ou elemento que possa, de qualquer forma, identificá-lo, será mantido em sigilo.

Proteção de informações pessoais na publicação dos resultados da investigação: Todo o material será utilizado unicamente para fins de pesquisa e será armazenado ao término do estudo.

As informações que o sr.(a) fornecer serão utilizadas para produzir um documento que será tornado público. Embora a informação bruta permanecerá confidencial, o pesquisador irá utilizar esta informação no trabalho apresentado para publicação.

### ASSINATURA

_____

## APÊNDICE 3 – LISTA DE PALAVRAS CONSIDERADAS *STOPWORDS* PARA UTILIZAÇÃO NA FERRAMENTA *ORANGE CANVAS*

de, a, o, que, e, do, da, em, um, para, é, com, não, uma, os, no, se, na, por, mais, as, dos, como, mas, foi, ao, ele, das, tem, à, seu, sua, ou, ser, quando, muito, há, nos, já, está, eu, também, só, pelo, pela, até, isso, ela, entre, era, depois, sem, mesmo, aos, ter, seus, quem, nas mexesse, eles, estão, você, tinha, foram, essa, num, nem, suas, meu, às, minha, têm, numa, pelos, elas, havia, seja, qual, será, nós, tenho, lhe, deles, essas, esses, pelas, este, fosse, dele, tu, te, vocês, vos, lhes, meus, minhas teu, tua, teus, tuas, nosso, nossa, nossos, nossas dela, delas, esta, estes, estas, aquele, aquela, aqueles, aquelas, isto, aquilo, estou, está estamos, estão, estive, esteve, estivemos, estiveram, estava, estávamos, estavam, estivera, estivéramos, esteja, estejamos, estejam, estivesse, estivéssemos, estivessem, estiver, estivermos, estiverem, hei, há, havemos, hão, houve, houvemos, houveram, houvera, houvéramos, haja, hajamos, hajam, houvesse, houvéssemos, houvessem, houver, houvermos, houverem, houverei, houverá, houveremos, houverão, houveria, houveríamos, houveriam, sou, somos, são, era, éramos, era, fui, foi, fomos, foram, fora, fôramos, seja, sejamos, sejam, fosse, fôssemos, fossem, for, formos, forem, serei, será, seremos, serão, seria, seríamos, seriam, ta, tenho, tem, temos, tém, tinha, tínhamos, tinham, tive, teve, tivemos, tiveram, tivera, tivéramos, tenha, tenhamos, tenham, tivesse, tivéssemos, tivessem, tiver, tivermos, tiverem, terei, terá, teremos, terão, teria, teríamos, teriam, vai, são, sao, pra, so, q, agora, agr, pro, ja, voces, vcs, #, ?, ., @, https, t, co, É, é.