

INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DE SÃO PAULO
CAMPUS CAMPINAS

Luiz Henrique Dória Santos

Mineração de Texto:
Análise de Sentimentos de Tweets referentes a Copa do Mundo de 2018.

Campinas
2018



Luiz Henrique Dória Santos

Mineração de Texto:

Análise de Sentimentos de Tweets referentes a Copa do Mundo de 2018.

Trabalho de Conclusão de Curso apresentado ao Curso de Tecnologia em Análise e Desenvolvimento de Sistemas do Instituto Federal de Educação, Ciência e Tecnologia de São Paulo – Campus Campinas, como requisito para a obtenção do grau de Tecnólogo em Análise e Desenvolvimento de Sistemas.

Orientador: Prof. Esp. Rafael da Silva Muniz.

Campinas

2018

Ficha catalográfica
Instituto Federal de São Paulo – Câmpus Campinas
Biblioteca
Danielle Sarmento – CRB 8/8669

S237m Santos, Luiz Henrique Doria
Mineração de texto: análise de sentimentos de tweets referentes a copa do mundo de 2018 / Luiz Henrique Doria Santos. – Campinas, SP: [s.n.], 2018.
50f. : il.

Orientador: Rafael da Silva Muniz.
Trabalho de Conclusão de Curso (graduação) – Instituto Federal de Educação, Ciência e Tecnologia de São Paulo Câmpus campinas. Curso de Tecnologia em Análise e Desenvolvimento de Sistemas, 2018.

1. Análise de sentimentos. 2. Mineração de textos. 3. Modelo de classificação. 4. Naive Bayes. 5. Python. I. Instituto Federal de Educação, Ciência e Tecnologia de São Paulo Câmpus Campinas. Curso de Tecnologia em Análise e Desenvolvimento de Sistemas. II. Título.

Luiz Henrique Dória Santos

Mineração de Texto:

Análise de Sentimentos de Tweets referentes a Copa do Mundo de 2018.

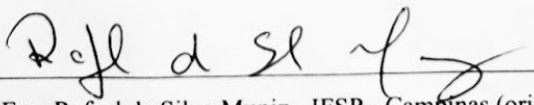
Trabalho de Conclusão de Curso apresentado ao Curso de Tecnologia em Análise e Desenvolvimento de Sistemas do Instituto Federal de Educação, Ciência e Tecnologia de São Paulo – Campus Campinas, como requisito para a obtenção do grau de Tecnólogo em Análise e Desenvolvimento de Sistemas.

Aprovado em: 30 de OUTUBRO de 2018.

BANCA EXAMINADORA

Profa. Ma. Joice Barbosa Mendes - IFSP - Campinas

Prof. Me. Fabio Feliciano de Oliveira - IFSP - Campinas


Prof. Esp. Rafael da Silva Muniz - IFSP - Campinas (orientador)

Dedico este trabalho à minha família, por tudo que sacrificaram em suas vidas ao me ajudarem a ser uma pessoa melhor.

AGRADECIMENTOS

Agradeço o meu orientador, Prof. Rafael Muniz, que dispôs seu tempo e esforço para me ajudar na elaboração deste documento e me orientar durante todo o desenvolvimento deste trabalho.

À minha família, que me motivou a seguir esse este caminho e não duvidaram de minha capacidade no decorrer este projeto.

Aos meus colegas de trabalho, que sempre me incentivaram em concluir, não só este projeto, mas também minha graduação.

Agradeço também a todos os meus professores e colegas de faculdade que se disponibilizaram em me ajudar no decorrer de todo o curso.

*“A paixão é o que te faz passar pelos momentos
mais difíceis que poderiam enfraquecer os homens
fortes, ou fazer você desistir..”*

(Neil deGrasse Tyson)

RESUMO

Nos últimos anos, a internet e, em específico, as redes sociais se tornaram plataformas onde os usuários publicam suas opiniões sobre produtos, serviços e eventos. Para as empresas e organizações, isso se torna um ótimo local para verificar o grau de aceitação de seus produtos e serviços. Porém, existe uma grande quantidade de dados disponíveis nas redes sociais que continuam crescendo, o que torna inviável a realização uma análise desses dados de forma manual. Com isso em mente, este trabalho teve como objetivo analisar comentários do *Twitter* referentes a Copa do Mundo de 2018, identificando a polaridade dos comentários, classificando-os como positivo, negativo ou neutro. Para atingir esse objetivo, foi utilizado os conceitos da Análise de Sentimentos, contidos na Mineração de Textos, e os métodos de classificação de dados Naive Bayes. Esses conceitos e métodos ajudaram na elaboração do modelo de classificação de dados que é utilizado para avaliar a polaridade dos comentários, de forma automática. Neste trabalho, é possível visualizar o nível de acurácia do modelo de classificação e a análise dos dados já classificados pelo modelo. Por fim, pode-se concluir que todos os conceitos e métodos utilizados, são de grande serventia para a análise de grandes quantidades de dados.

Palavras-chave: Análise de Sentimentos. Mineração de Textos. Modelo de Classificação. Naive Bayes. Python.

ABSTRACT

In recent years, the internet and, in particular, social networks have become platforms where users publish their opinions about products, services and events. For businesses and organizations, this becomes a great place to check the degree of acceptance of your products and services. However, there is a large amount of data available in social networks that continue to grow, which makes it impossible to carry out an analysis of this data manually. With this in mind, this work had as objective to analyze comments of Twitter referring to the 2018 World Cup, identifying the polarity of the comments, classifying them as positive, negative or neutral. To achieve this goal, the concepts of Sentiment Analysis, contained in the Text Mining, and Naive Bayes data classification methods were used. These concepts and methods helped in the elaboration of the data classification model that is used to evaluate the polarity of comments, automatically. In this work, it is possible to visualize the level of accuracy of the classification model and the analysis of the data already classified by the model. Finally, it can be concluded that all the concepts and methods used are of great use for the analysis of large amounts of data.

Keywords: Sentiment Analysis. Text Mining. Classification Model. Naive Bayes. Python.

LISTA DE TABELAS

Tabela 1 - Exemplo de classificação do Naive Bayes	38
Tabela 2 - Informações sobre a etapa de coleta de dados	40
Tabela 3 - Quantidade de <i>tweets</i> por base	41
Tabela 4 - Porcentagem de acurácia por polaridade	41
Tabela 5 - Total de <i>tweets</i> por polaridade	43
Tabela 6 - Porcentagem de acurácia por polaridade - Filho	43

LISTA DE FIGURAS

Figura 1 - Teorema de Bayes	22
Figura 2 - <i>Tweet</i> com <i>hashtag</i>	24
Figura 3 - <i>Tweet</i> sem <i>hashtag</i>	24
Figura 4 - Autenticação do usuário	28
Figura 5 - Código para utilizar a API	29
Figura 6 - Tabela da partida entre Brasil e Bélgica	30
Figura 7 - <i>Tweet</i> com comentário positivo	31
Figura 8 - <i>Tweet</i> com comentário negativo	31
Figura 9 - <i>Tweet</i> com comentário neutro	32
Figura 10 - Código para traduzir os comentários para Inglês	33
Figura 11 - Limite de caracteres por projeto e conta	33
Figura 12 - Método para remover elementos desnecessários	34
Figura 13 - Lista de elementos que foram removidos	35
Figura 14 - Lista de <i>stopwords</i>	36
Figura 15 - Método para remover as <i>stopwords</i> e <i>stemmers</i>	36
Figura 16 - Divisão da base em treinamento e teste	37
Figura 17 - Classificador Naive Bayes e acurácia	39
Figura 18 - Classificação dos <i>tweets</i> restantes	39
Figura 19 - Métricas das amostras por polaridade	42
Figura 20 - Classificação das polaridades por partida	44
Figura 21 - Classificação dos <i>tweets</i> do mês de junho	45
Figura 22 - Classificação dos <i>tweets</i> do mês de julho	46

LISTA DE ABREVIATURAS E SIGLAS

API	Application Programming Interface
AS	Análise de Sentimentos
CSV	Comma-Separated Values
FIFA	Fédération Internationale de Football Association
IE	Information Extraction
IR	Information Recovery
KDD	Knowledge Discovery in Databases
KDT	Knowledge Discovery from Text
NB	Naive Bayes
NTKL	Natural Language Toolkit
PLN	Processamento da Linguagem Natural
SGBD	Sistema de Gerenciamento de Banco de Dados

SUMÁRIO

1. INTRODUÇÃO	14
1.1 Justificativa	15
1.2 Objetivos	16
1.2.1 Objetivo Geral	16
1.2.2 Objetivo Específico	16
2. FUNDAMENTAÇÃO TEÓRICA	17
2.1 Análise de Sentimentos	17
2.2 Processamento da Linguagem Natural	18
2.3 Mineração Textual	20
2.4 Tweets e a Copa do Mundo de 2014 por José Filho	21
2.5 Classificador Naive Bayes	22
2.6 Twitter	24
2.7 Python	25
2.8 PostgreSQL	26
3. METODOLOGIA	27
3.1 Coleta de tweets	27
3.2 Pré-processamento dos dados	33
3.2.1 Google API Translate	33
3.2.2 Stopwords e Stemmer	36
3.3 Modelo de classificação	38
3.4 Classificação dos dados	40
4. RESULTADOS ALCANÇADOS	41
4.1 Coleta de Dados	41
4.2 Validação do Modelo de Classificação	42
4.4 Análise dos Sentimentos	45
5. CONCLUSÃO	48
REFERÊNCIAS	49

1. INTRODUÇÃO

Ao longo dos anos, as redes sociais vem se popularizando no mundo todo e vem servindo de ferramentas para a publicação de opiniões sobre eventos, produtos, organizações, dentre outros. Essas publicações podem ser positivas ou negativas, o que pode ser uma coisa benéfica ou prejudicial para os envolvidos no assunto.

Muitas empresas já perceberam a vantagem de conhecer as opiniões de seus usuários pela Internet, já que é possível verificar o nível de satisfação que os usuários têm sobre seus produtos ou serviços, porém, o volume de informações é grande que torna difícil de acompanhar todas as opiniões que estão sendo publicadas.

De acordo com Filho (2014), para preencher essa lacuna, a Mineração de Textos, focada em extrair padrões importantes em textos, juntamente com a Análise de Sentimentos, destinada a classificação dos mesmos, fornecem técnicas que auxiliam no estado aprofundado das opiniões em qualquer análise textual.

Este trabalho foi focado no tema Copa do Mundo de 2018, por ser um evento mundial e que aconteceu este ano de 2018. Assim como Filho (2014), cujo o trabalho de conclusão de curso foi avaliar os sentimentos dos usuários durante a Copa do Mundo de 2014, através do *Twitter*, é possível mapear a opinião dos usuários sobre este evento de escala mundial. Além das empresas conseguirem fazer uma análise de seus serviços e produtos durante este evento, é possível verificar o sentimento despertado em cada usuário sobre quaisquer assuntos ou temas, já que os métodos abordados neste trabalho servem como guia para o tratamento e a classificação das informações obtidas.

Por esses motivos, o objetivo deste trabalho foi desenvolver um software capaz de avaliar as opiniões dos usuários do *Twitter*, referentes à Copa do Mundo de 2018, classificando sua polaridade em positiva, negativa ou neutra. Com esse resultado foi possível refletir sobre o impacto que a Copa do Mundo teve sobre a população.

Este trabalho está dividido em seis seções, a primeira delas foi esta introdução com a contextualização do trabalho, na segunda seção é apresentada sua justificativa, na seção número três estão definidos o objetivo geral e os objetivos específicos. A

fundamentação teórica está localizada na quarta seção, onde é dividida em seis subseções, a metodologia se encontra na quinta seção e a conclusão do trabalho permanece na seção de número seis.

1.1 Justificativa

Devido a grande quantidade de textos publicados diariamente no *Twitter*, cerca de 500 milhões (STATS, 2018), a extração de *tweets* e a realização da análise de emoções desses textos se apresenta como uma tarefa difícil. Essa tarefa pode ser efetivada através da utilização de sistemas de informações, de forma automatizada.

Com a aplicabilidade da Mineração de textos, na extração de quaisquer informações pertinentes e a Análise de sentimentos, nas classificações das diversas opiniões, é possível realizar estudos em diversos campos do conhecimento (FILHO, 2014, p. 11). Os dados que saem como resultado destes estudos, fornecem informações importantes sobre a positividade ou a negatividade que estes comportamentos geram nos usuários em todas as partes do mundo.

Um exemplo de aplicação, utilizado nos dias atuais, é a verificação do nível de aceitação de um determinado produto, pelos usuários das redes sociais, em específico o *Twitter* (SANTOS, 2016, p. 46). Com isso, é possível verificar a quantidade de usuários que realmente gostaram, ou não, do produto e aplicar mudanças para melhorar a experiência do usuário.

Por conta dessas vantagens apresentadas, ao analisar as emoções e pela dificuldade em realização dessa atividade manual, foi proposto o desenvolvimento deste sistema a fim de auxiliar na extração, no tratamento e na análise das emoções destes textos.

1.2 Objetivos

1.2.1 Objetivo Geral

Desenvolver um software capaz de analisar os *tweets* referentes a Copa do Mundo de 2018, identificando a polaridade (positiva, negativa ou neutra) dos sentimentos das pessoas que comentaram sobre o assunto e observar a correlação entre os dados obtidos e os dados apresentados no trabalho de José Filho, referente a Copa do Mundo de 2014.

1.2.2 Objetivo Específico

- Coletar *tweets* relacionados à Copa do Mundo de 2018.
- Realizar o pré-processamento de todos os dados coletados.
- Dividir as amostras em duas partes, uma para treino e outra para teste.
- Gerar um modelo de classificação por meio da Análise de Sentimentos sobre as amostras de treinamento.
- Verificar o nível de acurácia do modelo.
- Analisar os dados obtidos sobre os sentimentos da população na Copa do Mundo de 2018, comparando com os dados da Copa de 2014.

2. FUNDAMENTAÇÃO TEÓRICA

Este capítulo apresenta as fundamentações teóricas necessárias para o aprofundamento do trabalho e está dividida em seis subseções, as duas primeiras estão destinadas à Análise de Sentimentos e ao Processamento da Linguagem Natural. Logo após é enfatizado sobre a mineração de texto, o classificador Naive Bayes e o *Twitter*, local de onde serão extraídos as opiniões dos usuários. Nas duas últimas subseções são discutidas sobre a linguagem de programação Python e o Sistema de Gerenciamento de banco de Dados PostgreSQL.

2.1 Análise de Sentimentos

As universidades e empresas estão focando, cada vez mais, na detecção de sentimentos nas redes sociais, pois “[...] é possível saber a todo momento, o que milhares, ou milhões de pessoas estão pensando, sentindo e expressando sobre assuntos de importância para quem estiver monitorando os usuários.” (MALHEIRO *et al.*, 2013, p. 2) e esse monitoramento pode ser inicialmente realizado pela Análise de Sentimentos.

Análise de Sentimentos (AS), também conhecida como análise de opinião, é definida como:

*[...] the field of study that analyzes people’s opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes [...]*¹ (LIU, 2012, p. 7).

¹ o campo de estudo que analisa as opiniões das pessoas, sentimentos, avaliações, estimativas, atitudes e emoções sobre entidades como produtos, serviços, organizações, indivíduos, problemas, eventos, tópicos e seus atributos (traduzido por este autor).

Ela pode ser utilizada para identificar a polaridade (positivo, negativo ou neutro) de um conteúdo publicado por um usuário nas redes sociais.

A AS é dividida em três etapas fundamentais para o levantamento de opiniões. A primeira etapa é composta pela coleta de dados, na segunda é realizada a classificação desses dados e na última etapa é feita a sumarização.

A coleta de dados visa apenas buscar na web conteúdos relacionados ao tema e arquivá-los para análise e classificação. A etapa de classificação pode ser realizada por meio de técnicas de aprendizagem de máquina, seleção de palavras ou análise sintática. E por fim, na sumarização de resultados, as classificações das diversas opiniões devem ser resumidas e sintetizadas, com o intuito de facilitar o seu entendimento sobre as mesmas (ARAUJO *et al.*, 2012, p. 96).

2.2 Processamento da Linguagem Natural

O Processamento da Linguagem Natural (PLN) é definido como um conjunto de técnicas computacionais que analisam e representam ocorrências de textos em vários níveis de análise linguística, tendo como objetivo aproximar o processamento de linguagem das máquinas ao dos seres humanos na realização de uma série de tarefas ou aplicações (ALLEN, 1995).

Natural Language Processing is a theoretically motivated range of computational techniques for analyzing and representing naturally occurring texts at one or more levels of linguistic analysis for the purpose of achieving human-like language

*processing for a range of tasks or applications.*² (LIDDY, 2001, p. 2).

De acordo com Gonzalez e Lima (2003), existem cinco níveis de processamento da linguagem natural, sendo elas: o fonético e fonológico, o morfológico, o sintático, o semântico e o pragmático. O PLN não necessariamente se comunica por meio de todos esses níveis, mas pode utilizar um ou mais níveis para o processamento da linguagem natural.

Ao longo dos anos, a PLN foi ganhando espaço com o avanço das tecnologias e a crescente necessidade de sistemas capazes de processarem informações advindas da população. Esses sistemas são, em sua maioria, ferramentas auxiliaadoras no processamento das diversas linguagens e no entendimento de seus significados.

São exemplos de aplicações de PLN, os correctores ortográficos, formatadores e hifenizadores, sistemas de extracção de informação, sumarizadores, tradutores automáticos e semiautomáticos, sistemas de pergunta-resposta, os sistemas interactivos, os conversores texto-fala, os reconhecedores de voz, entre outros (BRAGA *et al.*, 2008, p. 2).

As técnicas disponíveis no Processamento da Linguagem Natural serão utilizadas neste trabalho para identificar e remover todos os *tweets* que não apresentam quaisquer sentimentos ou polaridade, sobre o tema proposto.

² O Processamento da Linguagem Natural é um conjunto de técnicas computacionais para analisar e representar ocorrências naturais de texto em um ou mais níveis de análise linguística com o objetivo de se alcançar um processamento de linguagem similar ao humano para uma série de tarefas ou aplicações (traduzido por este autor).

2.3 Mineração Textual

A Mineração de texto é "[...] o processo de descobrir computacionalmente novas informações, previamente desconhecidas, pela extração automática de informação de diferentes recursos de texto." (HEARST, 2003, p. 1, tradução do autor). Ela é uma etapa pertencente à *Knowledge Discovery from Text*³ (KDT), que:

*[...] deals with the machine supported analysis of text. It uses techniques from information retrieval, information extraction as well as natural language processing (NLP) and connects them with the algorithms and methods of Knowledge Discovery in Databases (KDD), data mining, machine learning and statistics.*⁴
(HOTHOTH et al., 2005, p. 4).

A grande diferença entre a Mineração de Dados e a Mineração de Textos é que a primeira está focada em padrões de dados estruturados, como por exemplo banco de dados, já a segunda foca na análise de textos não estruturados, ligados à linguagem natural, como por exemplo as redes sociais.

A mineração de texto possui sete áreas de atuação, listadas a seguir:

1. Pesquisa e recuperação de informação (IR): armazenamento e recuperação de documentos de texto, incluindo ferramentas de pesquisa e palavras-chave;
2. Clusterização de documentos: agrupamento e categorização de termos, trechos, parágrafos ou documentos, utilizando métodos de clusterização de mineração de dados;

³ Descoberta de Conhecimento em Textos (traduzido por este autor).

⁴ lida com a análise de texto suportada pela máquina. Ele usa técnicas de recuperação de informações, extração de informações e processamento de linguagem natural (NLP) e as conecta com os algoritmos e métodos de Descoberta de Conhecimento em Bancos de Dados (KDD), mineração de dados, aprendizado de máquina e estatística (traduzido por este autor).

3. Classificação de documentos: agrupamento e categorização de trechos, parágrafos ou documentos usando métodos de classificação de mineração de dados, baseados em modelos treinados em exemplos pré-classificados;

4. Web mining: Mineração de dados e textos na Internet, com foco específico na escala e interconectividade da Web;

5. Extração de informação (IE): Identificação e extração de fatos e relacionamentos relevantes de textos não estruturados. Processo de tornar textos não estruturados e semiestruturados em dados estruturados;

6. Processamento de linguagem natural (PLN): Processamento de linguagem de baixo nível e compreensão de tarefas (por exemplo, tagging part-of-speech). Muitas vezes, usado como sinônimo para linguística computacional;

7. Extração de conceito: Agrupamento de palavras e frases em grupos semanticamente similares. (MINER *et al.*, 2012, p. 1009).

2.4 Tweets e a Copa do Mundo de 2014 por José Filho

Mineração de Texto: Análise de Sentimentos utilizando *Tweets* referentes à Copa do Mundo de 2014 foi um trabalho de conclusão de curso desenvolvido por José Adail Carvalho Filho, pela Universidade Federal do Ceará, no ano de 2014. Seu trabalho tinha como objetivo identificar o sentimento da população sobre a Copa do Mundo através das postagens no *Twitter* e correlacioná-los com os acontecimentos reais que aconteceram durante aquele evento.

Em seu trabalho, Filho (2014), justifica que devido a grande cobertura e interesse social da população, tanto pela popularidade do futebol brasileiro quanto pelas polêmicas e manifestações que rodeavam esse evento, torna-se interessante mapear a opinião dos usuários do *Twitter* sobre a Copa do Mundo.

O trabalho de Filho (2014) foi utilizado como base para o desenvolvimento e comparativo para este projeto. Os principais métodos expostos em seu trabalho, sendo eles: Mineração de Textos, Análise de Sentimentos, Processamento de Linguagem Natural e Classificador Naive Bayes, foram utilizados neste trabalho com a intenção de analisar se com os mesmos métodos é possível alcançar os mesmos resultados quatro anos depois.

2.5 Classificador Naive Bayes

Naive Bayes (NB) é uma técnica simples de classificação probabilística baseada no teorema de Bayes. Esse teorema, “[...] supõe que o efeito do valor de um atributo em uma determinada classe seja independente dos valores dos outros atributos.” (LEUNG, 2007, p. 3, tradução do autor). A figura 1 abaixo, ilustra o teorema de Bayes.

Figura 1: Teorema de Bayes.

$$p(c|E) = \frac{p(E|c)p(c)}{p(E)}$$

Fonte: (ZHANG, 2004).

De acordo com Filho (2014):

E representa um evento que ocorreu previamente e c um evento que depende de E , para que seja calculada a probabilidade de c ocorrer dado o evento E , o algoritmo deverá contar o número de casos em que c e E ocorrem

juntos e dividir pelo número de casos em que *E* ocorre sozinho (FILHO, 2014, p. 17).

Por ser um classificador probabilístico simples de ser utilizado e ser “[...] considerado um dos mais eficientes em questões relacionadas com processamento e precisão na classificação de novas amostras.” (GOMES, 2013, p.23), o classificador Naive Bayes foi utilizado neste trabalho para classificar a polaridade dos *tweets* coletados. Outro motivo para utilizar este classificador é devido a utilização do mesmo no trabalho de Filho (2014). Decidiu-se utilizar o mesmo método de classificação para avaliar se, com o mesmo método, é possível adquirir os mesmo resultados, quatro anos depois.

2.6 Twitter

O *Twitter* é uma rede social, que permite aos seus usuário publicarem informações e ideias em tempo real. De acordo com Stats (2018), são publicados aproximadamente 500 milhões de *tweets* por dia. Cerca de 316 milhões de usuários estão publicando ativamente todos os meses, o que mostra a relevância dessa rede sociais na amplitude de seu alcance.

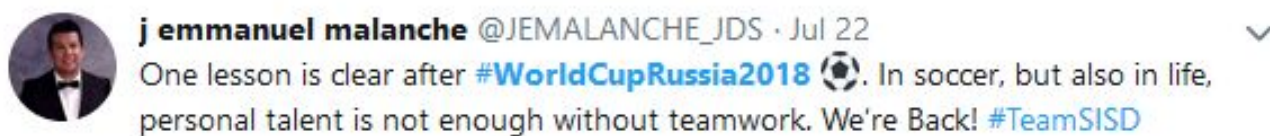
Desde o começo do *Twitter* até Setembro de 2017, os *tweets* eram limitados a 140 caracteres, porém, como Rosen (2017) descreve, ela acabava chegando no limite muito rapidamente e acabava por ter que alterar a frase para que ficasse dentro dos 140 caracteres.

Com isso, o número total de caracteres disponíveis foi aumentado para 280. Contudo, como afirma Rosen (2017), alguns idiomas como o japonês, coreano e chinês, transmitem o dobro da quantidade de informações, com menos caracteres, o que os tornam os únicos que não desfrutaram dessa atualização.

Além de textos, os usuário podem enviar links, fotos, *gifs*⁵ animados e vídeos, deixando seus *tweets* mais dinâmicos, que não serão considerados neste trabalho.

Uma ferramenta de categorização muito utilizada pelos usuário é a *hashtag*, representada pelo # seguido por uma ou mais palavras chaves, são utilizadas para marcar o assunto do *tweet*. Esta ferramenta será de grande importância para este trabalho, já que os *tweets* sobre a Copa do Mundo de 2018 serão extraídos pela busca de *hashtags* relacionadas ao tema. A figura 2 apresenta um *tweet* utilizando *hashtags* e a figura 3 não.

Figura 2: *Tweet* com *hashtag*.



Fonte: *Twitter*.

Figura 3: *Tweet* sem *hashtag*.



Fonte: *Twitter*.

2.7 Python

Python é uma linguagem de programação de alto nível e de propósito geral que proporciona um aprendizado rápido e uma manutenção simples. Por ter muitas bibliotecas prontas para analisar dados científicos, "[...] A linguagem de programação

⁵ Graphics Interchange Format (formato de intercâmbio de gráficos).

Python está se estabelecendo como uma das linguagens mais populares para computação científica" (PEDREGOSA *et al.*, 2011, p. 2826, tradução do autor).

"*Thanks to its high-level interactive nature and its maturing ecosystem of scientific libraries, it is an appealing choice for algorithmic development and exploratory data analysis*"⁶ (MILLMAN; AIVAZIS, 2011, p. 9). As bibliotecas mais utilizadas na Análise de Sentimentos são o *Natural Language Toolkit* (NLTK), o SKLearn e o Matplotlib.

O NLTK é responsável por processar as frases recolhidas no *Twitter* e classificá-los de acordo com os sentimentos encontrados, o SKLearn é comumente utilizado para a manipulação de vetores N-dimensionais e para cálculos matemáticos e o Matplotlib é utilizado para mostrar diferentes gráficos das informações coletadas. Essas são as três bibliotecas mais importantes, em termos de tratamento de dados científicos, e serão as mais utilizadas neste trabalho.

2.8 PostgreSQL

O PostgreSQL é um Sistema de Gerenciamento de Banco de Dados (SGBD) que armazena e gerencia todos dados coletados de uma aplicação. De acordo com PostgreSQL (2018), o sistema de banco de dados de código aberto, PostgreSQL, usa a linguagem SQL combinada com outros recursos para armazenar e dimensionar os diferentes tipos de dados.

O PostgreSQL está disponível, atualmente, para vários tipos de Sistemas Operacionais, o que o torna uma excelente ferramenta. Para este trabalho o PostgreSQL será utilizado para armazenar e gerenciar todos os *tweets* obtidos pela plataforma *Twitter*, sobre o tema proposto. Esses *tweets* foram obtidos no decorrer da Copa do Mundo de 2018 e inseridos automaticamente no banco de dados para que, posteriormente, pudessem ser utilizados na classificação dos sentimentos.

⁶ Graças à sua natureza interativa de alto nível e suas bibliotecas específicas para a ciência, é uma opção atraente para o desenvolvimento de algorítmico e a análise de dados (traduzido por este autor).

3. METODOLOGIA

Nesta seção estão dispostas, de forma detalhada, cada etapa percorrida no desenvolvimento do sistema, tomando como base os objetivos específicos, dispostos na subseção 3.2 deste trabalho, com o intuito de atingir o objetivo geral estabelecido na subseção 3.1.

Nesta seção serão tratados sobre como foram coletados os dados do *Twitter*, levando em consideração todas as ferramentas necessárias e suas limitações. Também será tratado como foi feito o pré-processamento desses dados, de como foi realizado o modelo de classificação e de como os dados foram classificados de acordo com o modelo.

3.1 Coleta de *tweets*

Antes mesmo de coletar os *tweets* é necessário criar uma conta no *Twitter* e habilitar o modo desenvolvedor para ter acesso às APIs. Após habilitado a conta como desenvolvedor, é gerado uma *consumer key* e uma *consumer secret* que representam a chave de acesso do usuário cadastrado, porém, só com essas chaves não é o suficiente para fazer as requisições pela API, é necessário adquirir o *access token* e o *access token secret* para realizarem essas requisições.

É importante ressaltar que a API do *Twitter* tem certas limitações, com relação às conexões. O próprio Twitter (2018) informa que se o cliente tentar estabelecer várias conexões com a mesma credencial, a conexão mais antiga é terminada. Neste trabalho foi possível verificar que duas conexões, executadas simultaneamente, são o máximo permitido para que não tenha nenhum tipo de interferência.

Tendo a base de quantas conexões são possíveis realizar sem ter nenhum prejuízo, foi realizado um estudo de quantos jogos seriam efetuados por dia e quais seriam os

horários deles, para identificar quantas conexões são necessárias em todos os dias. Na fase de grupos os jogos estavam distribuídos em três partidas por dia, sendo uma de manhã, uma no horário do almoço e outra de tarde. Para esta fase foi constatado que duas conexões são o suficiente, já que os dados a serem coletados são sobre cada partida e sobre a Copa do Mundo como um todo, com isso, uma conexão iria coletar os dados da partida enquanto a outra fica coletando sobre a Copa.

Nas oitavas de final, foi constatado que três conexões seriam necessárias, porque as partidas seriam realizadas simultaneamente, tanto no período da manhã quanto no período da tarde. Nesta ocasião, seria necessário criar uma outra conta no *Twitter* para que não houvesse qualquer restrição na coleta dos dados. Já nas quartas de final, nas semifinais e a final, o número de conexões poderia voltar para duas, já que as partidas seriam disputadas em horários diferentes.

Sabendo quantas conexões e quantas contas seriam necessárias para o projeto, foi imprescindível escolher qual seria o método de coleta dos dados, se seria feita por palavras chaves dentro do texto ou pelas *hashtags*. Foi optado por utilizar as *hashtags*, já que é fácil de identificar um tema ou um conteúdo em específico. Se fosse utilizado o método de palavras chaves no texto, teria que se tomar um certo cuidado para não coletar dados que não estejam relacionados com o tema proposto, isso poderia acarretar na coleta de dados que não teriam qualquer ligação com as partidas entre as seleções.

Após determinado o método de coleta, tornou-se necessário avaliar quais seriam as *hashtags* mais utilizadas para identificar cada partida e também a Copa do Mundo como um todo. Felizmente, a grande maioria dos usuários estavam utilizando as *hashtags* disponibilizadas pela *Fédération Internationale de Football Association* (FIFA). No começo da Copa do Mundo, a FIFA definiu a *hashtag* *#WorldCupRussia2018* como sendo a *hashtag* oficial da Copa do Mundo e no começo de cada dia ia disponibilizando as *hashtags* de cada uma das partida.

A FIFA estava seguindo um padrão para definir como seriam as *hashtags* relacionadas a cada time e a cada partida. Esse padrão se dava pelas iniciais de cada time, compostas por três letras maiúsculas. Em cada partida as iniciais dos times eram

concatenadas, de forma a comporem uma única palavra. Por exemplo, a seleção do Brasil tinha como *hashtag* relacionada a ela, a *#BRA* e a Bélgica tinha como *hashtag* *#BEL*, quando as duas seleções se enfrentaram a *hashtag* destinada para a partida foi *#BRABEL*.

Com essa padronização é possível prever quais seriam todas as *hashtags* de todos os times e de todas as partidas que ocorreriam na fase de grupos, porém, nas fases subsequentes, não é possível prever quais seleções conseguiriam passar, então adotamos o método de analisar quais *hashtags* utilizaríamos, no dia em que as partidas iriam ser realizadas.

Após verificar tudo que precisávamos utilizar para coletar os *tweets*, a atenção foi voltada para a implementação e logo no começo foi constatado que já existe uma biblioteca pronta em Python que consegue coletar os *tweets*. Essa biblioteca se chama *tweepy* e ela consegue fornecer uma vasta aplicabilidade quando se trata do *Twitter*. Para utilizarmos essa biblioteca é necessário instalá-la, por meio do *pip* do próprio Python, como mostra o comando abaixo.

\$ pip install tweepy

Ao finalizar a instalação do *tweepy*, já é possível utilizá-lo no código para coletar os dados, porém, é necessário utilizar as chaves de acesso geradas ao criar a conta, para que o *Twitter* possa autenticar o seu usuário. Um exemplo do código em Python de como deve ser realizado essa autenticação, está disposto na figura 4 abaixo.

Figura 4: Autenticação do usuário.

```
1  import tweepy
2  from tweepy import OAuthHandler
3
4  consumer_key = 'YOUR-CONSUMER-KEY'
5  consumer_secret = 'YOUR-CONSUMER-SECRET'
6  access_token = 'YOUR-ACCESS-TOKEN'
7  access_secret = 'YOUR-ACCESS-SECRET'
8
9  auth = OAuthHandler(consumer_key, consumer_secret)
10 auth.set_access_token(access_token, access_secret)
11
12 api = tweepy.API(auth)
```

Fonte: (BONZANINI, 2015).

O *Twitter* disponibiliza várias APIs de coleta de dados, mas a que foi utilizada para este trabalho foi a de *StreamListener*. Esta API faz com que a conexão fique aberta para coletar os dados de um determinado evento, possibilitando que todos os *tweets* enviados por uma determinada *hashtag* sejam coletados enquanto o código estiver rodando. Esta API é muito útil, já que queríamos coletar os dados pela *hashtag* e enquanto as partidas estiverem acontecendo. O código que usufrui dessa API está disposto na figura abaixo.

Figura 5: Código para utilizar a API.

```
1  from tweepy import Stream
2  from tweepy.streaming import StreamListener
3
4  class MyListener(StreamListener):
5
6      def on_data(self, data):
7          try:
8              with open('python.json', 'a') as f:
9                  f.write(data)
10                 return True
11             except BaseException as e:
12                 print("Error on_data: %s" % str(e))
13                 return True
14
15         def on_error(self, status):
16             print(status)
17             return True
18
19     twitter_stream = Stream(auth, MyListener())
20     twitter_stream.filter(track=['#python'])
```

Fonte: (BONZANINI, 2015).

O código da figura 5 coleta todos os *tweets*, relacionados a *hashtag* informado na linha 20, em tempo real. Esses *tweets* são salvos no arquivo desejado, previamente informado como mostra na linha 8, no formato JSON, porém, utilizamos o banco de dados PostgreSQL para salvar esses dados. O dado retornado da API vem com várias informações, não só do texto digitado, mas também todos os dados do usuário que enviou o *tweet* e se o texto for um *retweet*, é informado também o texto proveniente do *retweet* e os dados do usuário que escreveu esse texto.

Como a API nos disponibiliza várias informações, optamos por coletar somente o nome do usuário, sua localização, o texto digitado, o idioma do usuário e o idioma do texto, porque verificamos que o resto das informações não são essenciais para este projeto (Figura 6). Esses dados são salvos em tabelas no banco de dados, de acordo com cada partida, então para cada partida é criado uma tabela no banco de dados e inseridos os *tweets* referentes a essa partida.

Figura 6: Tabela da partida entre Brasil e Bélgica.

id	user_name	user_location	tweet_text	user_lang	tweet_lang
1	Jota Pê	Marituba	Começou o aquecimento aqui em casa!! #World...	pt	pt
2	Kyama	Somewhere	Brazil vs Belgium is the match where the VAR ca...	en	en
3	Gabriel 🌟	Rio de Janeiro, Bra...	O que será que tem dentro da necessaire doura...	pt	pt
4	Ro 🌟	Montevideo, Urug...	#BRA es el único equipo tercer-mundista, latino...	es	es
5	Meghan Kavanaugh	Colorado	RT @BelgiumMFA: Hey, @BelRedDevils, just to le...	en	en
6	Gothboi BEPT	Paris, France	J'espère que le Brésil va gagner 🌟🌟🌟🌟🌟...	en	fr
7	Junny'ereleston 🌟	São Bernardo do ...	Alisso 6 LETRAS	pt	pt
8	elly	lostincmla	A MOÇA BEIJANDO O ENTREVISTADOR KKKKKKK...	pt	pt
9	RafaPT	None	RT @FootballBattles: RT for Lukaku🇧🇪	pt	en
10	ya boy sergioes💧13🏆	Upside down	RT @FootballBattles: RT for Neymar🇧🇷	nl	en
11	100% Neymar	Barão de Cocais, B...	RT @softgxrrl: Enquanto isso na torcida brasileira	pt	pt
12	Yasmin VEM HEXA 🇧🇷	no seu coração	eu acho que vai para os pênalti hein #BRABEL h...	pt	pt
13	Dan	Rio de Janeiro, Bra...	RT @KelvinLopes22: Chegando pra mais um dia ...	pt	pt
14	🔥Souzinha🔥GO BRA...	Marte	RT @IsaacNervoso: ELIMINADAS:	pt	pt
15	Reece BESE	None	RT @CelticFC: 🇬🇪 @kierantierney1's prediction f...	en	en

Fonte: criado por este autor.

Em média, foi constatado que cada partida foi coletado em torno de 124.877 *tweets* e em algumas partidas, como Portugal e Espanha e também a final entre a França e a Croácia, chegou em torno dos 300.000 *tweets*.

No final desta etapa de coleta de todos os dados das partidas da Copa do Mundo e também da *hashtag* sobre a copa do mundo, foi feito o levantamento do total de *tweets*

coletados e foi obtido 7.492.664 tweets. Desse total, todas as partidas somaram 6.808.779 tweets e a hashtag #WorldCupRussia2018 chegou a 683.885. As soma das cinco partidas que o Brasil participou chegaram a 837.932 tweets.

Com essa quantidade de dados é possível realizar a análise de sentimentos, porém, só após o pré-processamento de dados pode-se verificar quantos tweets, exatamente serão utilizados, já que nem todos os tweets coletados demonstram algum tipo de sentimento sobre a partida. Existem alguns tweets que simplesmente apresentam os placares durante a partida e também alguns dados estatísticos relacionados aos jogadores.

As três Figuras abaixo, exemplificam comentários positivo, negativo e neutro. A Figura 7 mostra um tweet positivo, a Figura 8 mostra um tweet negativo e a Figura 9 mostra um tweet neutro.

Figura 7: Tweet com comentário positivo.



Fonte: Twitter.

Figura 8: Tweet com comentário negativo.



Fonte: Twitter.

Figura 9: *Tweet* com comentário neutro.



Fonte: *Twitter*.

3.2 Pré-processamento dos dados

Para realizar o pré-processamento dos dados, foi necessário traduzir os *tweets* para um único idioma, já que foram coletados dados de diferentes idiomas. Com os *tweets* traduzido, foi necessário remover palavras e conteúdos que não serão utilizados na classificação desses dados, deixando que apenas o necessário seja utilizado para gerar o modelo de classificação. Nas subseções a seguir estão detalhados os processos de tradução para um único idioma e como foi feita a remoção de palavras desnecessárias.

3.2.1 *Google API Translate*

Por conta dos diferentes tipos de idiomas encontrados nos dados coletados, foi necessário traduzi-los para um único idioma. Ao pesquisar sobre métodos ou APIs que realizassem essa tradução automaticamente, foi constatado que a API do *Google* poderia ser de grande ajuda. Na documentação do *Google Translate API* (CLOUD, 2018), foi possível verificar que um dos métodos disponíveis, identifica automaticamente o idioma do texto e traduz para para outro idioma de sua escolha. O código que mostra esse método está na figura 10, logo abaixo.

Figura 10: Código para traduzir os comentários para Inglês.

```
def translate_text(text, target='en'):
    translate_client = translate.Client.from_service_account_json('apikey.json')
    result = translate_client.translate(text, target_language=target)
    tmp = []
    for row in result:
        tmp.append(row['translatedText'])
    return tmp
```

Fonte: criado por este autor.

Decidiu-se utilizar o Inglês como linguagem padrão para a análise, porque algumas bibliotecas do NLTK já são otimizadas para essa linguagem. Infelizmente o *Google* tem certas limitações com relação ao uso da sua API de tradução. Na documentação da API é possível verificar que ao passar de 1 milhão de caracteres em 100 segundos a conexão é encerrada (Figura 11). Para evitar chegar no limite, foi necessário adicionar uma verificação no código, para que após uma certa quantidade de dados a aplicação aguarde alguns segundos antes de continuar.

Figura 11: Limite de caracteres por projeto e conta.

Content Quota	Default	Maximum	Duration	Applies To
Characters per day	1 billion	unlimited	day	project
Characters per 100 seconds per project	1,000,000 characters	10,000,000 characters	100 seconds	project
Characters per 100 seconds per project per user	100,000 characters	10,000,000 characters	100 seconds	user and project

Fonte: (GOOGLE, 2018).

Outro problema envolve o limite de caracteres gratuitos. De acordo com a documentação, o limite de caracteres disponíveis gratuitamente é de 500 mil, após ultrapassar esse limite é cobrado 80 dólares a cada 5 milhões de caracteres traduzidos. Infelizmente, 500 mil não cobre todos os dados que necessitam tradução, então foi decidido fazer a Análise de Sentimentos somente das partidas em que o Brasil participava.

Dos 7 milhões de *tweets* coletados, reduziu-se para 823.072, porém, ainda era muitos dados para serem traduzidos. Um grande ajuda, foi que ao criar uma conta no *Google Cloud* o usuário recebe uma quantia inicial para gastar nos serviços oferecidos. Este valor começou com 300 reais, mas posteriormente, com algumas atualizações, subiu para 1.200 reais.

Para tirar o melhor proveito do benefício que a *Google* disponibilizou, foi necessário remover algumas informações inúteis para a análise, como os *emoticons*, *links* de páginas web, as *hashtags*, etc. Essa remoção aconteceu, automaticamente, via código de programação, com ela, foi possível reduzir o número de caracteres a serem traduzidos e manter somente os textos que eram necessários para fazer a tradução. O código correspondente a remoção e a lista de caracteres que foram removidos estão presentes na Figura 12 e 13, respectivamente.

Figura 12: Método para remover elementos desnecessários.

```
def preprocess(s, lowercase=False):
    tokens = tokenize(s)
    if lowercase:
        tokens = [token if emoticon_re.search(token) else token.lower() for token in tokens if token not in stop]
    return tokens
```

Fonte: criado por este autor.

Figura 13: Lista de elementos que foram removidos.

```
emoticons_str = r"""
(?:
    [:=;] # Eyes
    [oO\~]? # Nose (optional)
    [D\)\]\(\)/\OpP] # Mouth
)"""

regex_str = [
    emoticons_str,
    r'<[^>]+>', # HTML tags
    r'(?:@[\w_]+)', # @-mentions
    r'(?:\#+[\w_]+[\w\'\_\-]*[\w_]+)', # hash-tags
    r'http[s]?://(?:[a-z]|[0-9]|[\~_@.amp;+]|[!*\(\),]|(?:%[0-9a-f][0-9a-f]))+', # URLs

    r'(?:(?:\d+,)?+(?:\.\d+)?)', # numbers
    r'(?:[a-z][a-z\_\-]+[a-z])', # words with - and _
    r'(?:[\w_]+)', # other words
    r'(?:\S)' # anything else
]
```

Fonte: criado por este autor.

Com essa ajuda financeira da *Google*, a remoção dos elementos desnecessários e a redução do escopo para somente as partidas do Brasil, foi possível realizar a tradução dos diferentes idiomas para o Inglês. Mas só com a criação de mais três contas no *Google Cloud* foi possível traduzir tudo, porque a quantidade de caracteres que precisavam ser traduzidos, ultrapassava o limite disponível e necessitava muito mais do que 1.200 reais.

3.2.2 *Stopwords* e *Stemmer*

Stopwords são palavras, comumente usadas, que normalmente são ignoradas em pesquisas e consultas de dados. Essas palavras, no âmbito da Análise de Sentimentos, não acrescentaram nenhum tipo de sentimento ao texto e por isso elas foram removidas. Na figura 14 está a lista das *stopwords* em Inglês que foram removidas de todos os dados.

Figura 14: Lista das *stopwords*.

```
['i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you', "you're",  
"you've", "you'll", "you'd", 'your', 'yours', 'yourself', 'yourselves', 'he',  
'him', 'his', 'himself', 'she', "she's", 'her', 'hers', 'herself', 'it', "it's",  
'its', 'itself', 'they', 'them', 'their', 'theirs', 'themselves', 'what', 'which',  
'who', 'whom', 'this', 'that', "that'll", 'these', 'those', 'am', 'is', 'are', 'was',  
'were', 'be', 'been', 'being', 'have', 'has', 'had', 'having', 'do', 'does', 'did',  
'doing', 'a', 'an', 'the', 'and', 'but', 'if', 'or', 'because', 'as', 'until', 'while',  
'of', 'at', 'by', 'for', 'with', 'about', 'against', 'between', 'into', 'through',  
'during', 'before', 'after', 'above', 'below', 'to', 'from', 'up', 'down', 'in', 'out',  
'on', 'off', 'over', 'under', 'again', 'further', 'then', 'once', 'here', 'there', 'when',  
'where', 'why', 'how', 'all', 'any', 'both', 'each', 'few', 'more', 'most', 'other', 'some',  
'such', 'no', 'nor', 'not', 'only', 'own', 'same', 'so', 'than', 'too', 'very', 's', 't',  
'can', 'will', 'just', 'don', "don't", 'should', "should've", 'now', 'd', 'll', 'm', 'o',  
're', 've', 'y', 'ain', 'aren', "aren't", 'couldn', "couldn't", 'didn', "didn't", 'doesn',  
"doesn't", 'hadn', "hadn't", 'hasn', "hasn't", 'haven', "haven't", 'isn', "isn't", 'ma',  
'mightn', "mightn't", 'mustn', "mustn't", 'needn', "needn't", 'shan', "shan't", 'shouldn',  
"shouldn't", 'wasn', "wasn't", 'weren', "weren't", 'won', "won't", 'wouldn', "wouldn't"]
```

Fonte: criado por este autor.

Além de remover as *stopwords*, precisava-se também remover os radicais, para que as palavras como *love* (amor) e *loved* (amado) fossem identificados como sendo a mesma palavra. Essa remoção dos radicais é chamada de *stemmer*. A figura 15 mostra o código que remove as *stopwords* e o *stemmer* de cada *tweet*.

Figura 15: Método para remover as *stopwords* e *stemmers*.

```
def removeStemmer(list):  
    ENstopwords = stopwords.words('english') + punctuation + regex_str  
    stemmer = stem.PorterStemmer()  
  
    phrase = []  
    for row in list:  
        low_text = row.lower()  
        tmp = re.sub(r'[\^\w]', ' ', low_text)  
        result = [str(stemmer.stem(element)) for element in tmp.split() if element not in ENstopwords]  
        phrase.append((result, ''))  
  
    return phrase
```

Fonte: criado por este autor.

Ao final desse processo, todos os 837.932 *tweets* sobre as partidas do Brasil estavam quase prontos para serem utilizados na criação do modelo de classificação dos dados. O único elemento que estava faltando foi a base dos dados já classificados como positivo, negativo e neutro. Sem essa base, não seria, para este trabalho, possível gerar um modelo de classificação. Então, como etapa final do processo de pré-processamento, o autor deste trabalho classificou, manualmente, a polaridade de 2 mil *tweets* para serem utilizados no modelo de classificação.

3.3 Modelo de classificação

Para começar a elaboração do modelo, foi necessário dividir a base de dados já classificada em uma base de treinamento e outra de teste, assim é possível gerar o modelo de classificação com base nos dados de treinamento e, posteriormente, validados com os dados de teste.

Para fazer essa divisão dos dados, foi utilizado a biblioteca *sklearn*. Essa biblioteca tem alguns métodos em Python que já realizam essa divisão automaticamente e aleatoriamente. O código que faz essa separação dos dados está na figura 16, logo abaixo.

Figura 16: Divisão da base em treinamento e teste.

```
from sklearn.model_selection import train_test_split

base_train, base_test = train_test_split(base, test_size=0.33, random_state=14)
complete_base_train = nltk.classify.apply_features(extract_words, base_train)
complete_base_test = nltk.classify.apply_features(extract_words, base_test)
```

Fonte: criado por este autor.

Com a função *train_test_split*, mostrada na figura 16, é possível determinar a porcentagem de dados que irão para a base de teste, representado pelo *test_size*, assim, poderia se limitar a quantidade de dados que seriam usados para validar o modelo, já que

no âmbito de *Machine Learning*, dependendo da quantidade dos dados, a recomendação é de dividi-los em 70% para treinamento e 30% para teste. Neste trabalho decidiu-se utilizar 33% de nossa base para teste e 67% para treinamento. Então dos 2 mil dados já classificados, 660 foram para teste e 1.340 foram para treinamento.

Com a base de treinamento pronta, foi utilizado uma função do NLTK que já faz todo o processo do Naive Bayes. Essa biblioteca gera uma tabela com todas as palavras contidas na base de dados e contabiliza quantas vezes cada palavra está presente nos dados e quantas vezes cada palavra está presente nas classes que escolhemos (positivo, negativo, neutro). A tabela 1 abaixo é um exemplo de como o classificador Naive Bayes funciona.

Sentimentos	love	hate	good	bad	like	soccer	game
positivo	7/10	0/8	8/9	0/11	5/10	2/7	2/10
negativo	0/10	8/8	0/9	9/11	1/10	0/7	1/10
neutro	3/10	0/8	1/9	2/11	4/10	5/7	7/10

Tabela 1: Exemplo de classificação do Naive Bayes (criado por este autor).

Nessa tabela, pode-se verificar que a palavra *love* está presente em 10 frases distintas e dessas 10, 7 são frases positivas e 3 são consideradas neutras. Assim como a palavra *hate*, que está presente em 8 frases e dessas 8, todas tem um sentimento negativo. Dessa forma o NB gera um modelo de classificação probabilístico com base nessa tabela que ele mesmo cria.

Neste projeto, ao utilizar o método de classificação na base de treinamento, todas as palavras presentes nela foram contabilizadas, como no exemplo da Tabela 1 e dessa forma foi gerado o modelo de classificação que, posteriormente, foi utilizado na base de teste para verificar o nível de acurácia que nosso modelo teve. O código de classificação da base de treinamento para gerar o modelo de classificação e o código para a verificação da acurácia na base de teste, estão na Figura 17.

Figura 17: Classificador Naive Bayes e acurácia.

```
classified = nltk.NaiveBayesClassifier.train(complete_base_train)
print(nltk.classify.accuracy(classified, complete_base_test))
```

Fonte: criado por este autor.

3.4 Classificação dos dados

Com o modelo de classificação criado sobre a base de treinamento e com um grande aproveitamento na acurácia do modelo sobre a base de teste, foi possível classificar, automaticamente, o restante dos 823.072 *tweets* referentes às partidas do Brasil. O código utilizado para classificar o restante dos *tweets* se encontra na Figura 18.

Figura 18: Classificação dos *tweets* restantes.

```
def classifyPhrase(list, classified):
    template = []
    for (phrase, emotion) in list:
        result = extract_words(phrase)
        template.append((phrase, classified.classify(result)))
```

Fonte: criado por este autor.

Após, aproximadamente 11 horas ininterruptas de processamento. Todos os dados foram processados e armazenados em um arquivo CSV, para que fossem utilizados na elaboração dos resultados apresentados na Seção 6.

4. RESULTADOS ALCANÇADOS

4.1 Coleta de Dados

Ao final da Copa do Mundo de 2018, foram coletados 7.492.664 *tweets*. Sendo que desse 7 milhões, 837.932 foram das partidas do Brasil. A tabela 2 abaixo mostra algumas informações relacionadas a etapa de coleta de dados.

Descrição	Quantidade de <i>Tweets</i>
Todas as partidas	6.808.779
Todas as partidas do Brasil	837.932
Sobre a Copa do Mundo	683.885
Média por partida	124.877
Total coletado	7.492.664

Tabela 2: Informações sobre a etapa de coleta de dados (criado por este autor).

As partidas mais comentadas foram, o jogo entre a Nigéria e a Islândia, com 386.855 *tweets*, a final entre França e Croácia, com 325.175 *tweets* e a partida entre Portugal e Espanha, com 260.574 *tweets*. Entre as 5 partidas do Brasil, a mais comentada foi do Brasil e Suíça, com 220.460 *tweets*.

Em comparação com o trabalho realizado por Filho (2014), este trabalho coletou 5.363.802 *tweets* a mais sobre a Copa do Mundo, porém, devido às limitações apresentadas anteriormente, foi decidido utilizar somente os *tweets* referentes às partidas do Brasil. Por essa limitação, o trabalho de Filho teve uma base maior para realizar a Análise de Sentimentos, essa base foi de 2.128.862 *tweets*.

De todos os dados coletados sobre as partidas do Brasil, este autor classificou manualmente 2 mil comentários em positivo, negativo e neutro. Desses 2 mil, 33% foram

selecionados para a base de teste e os 67% restantes, se tornaram a base de treinamento. A tabela 3 mostra a quantidade de *tweets* dividido entre a base de treinamento e a base de teste.

Base	Quantidade de <i>tweets</i>
Treinamento	1340
Teste	660

Tabela 3: Quantidade de *tweets* por base (criado por este autor).

No trabalho de Filho (2014) a quantidade de amostras utilizada na base de treinamento foi de 3.250 e sua base de teste foi de 312. Em comparação com a bases deste trabalho, a quantidade de amostras utilizada foi inferior a quantidade utilizada por Filho, o que pode ter afetado o nível de precisão que o modelo de classificação teve, já que nossa base de treinamento foi pequena.

4.2 Validação do Modelo de Classificação

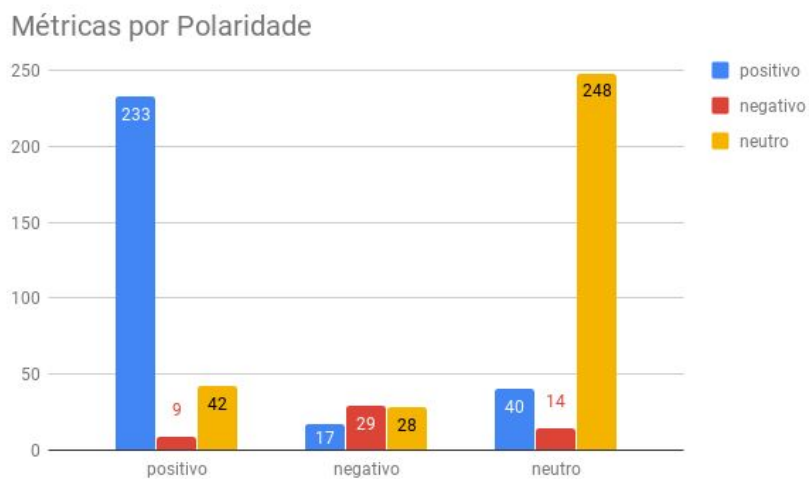
O modelo de classificação que obteve a maior acurácia com a base de treinamento, foi o que alcançou 77,27% de precisão. Com essa porcentagem, nosso modelo inicial conseguiu prever, com 77,27% de acurácia, a polaridade dos sentimentos das amostras contidas na base de teste.

Polaridade	Porcentagem
Positivo	82%
Negativo	39%
Neutro	82%

Tabela 4: Porcentagem de acurácia por polaridade (criado por este autor).

A Tabela 4, exibe a porcentagem alcançada para cada uma das polaridades, na base de teste. O modelo teve um precisão de 82% ao classificar, corretamente, as amostras consideradas positivas e neutras, mas não teve um bom aproveitamento ao classificar as amostras com polaridade negativa.

Figura 19: Métricas das amostras por polaridade.



Fonte: criado por este autor.

A Figura 19 mostra a quantidade de *tweets* que o modelo classificou, corretamente, em cada uma das polaridades e também a quantidade de *tweets* que classificou incorretamente. Dos 284 *tweets* já classificados como positivo (Tabela 5), o modelo conseguiu classificar corretamente 233 deles. 9 *tweets* negativos e 42 *tweets* neutros foram classificados incorretamente.

Polaridade	Total de <i>tweets</i>
Positivo	284
Negativo	74
Neutro	302

Tabela 5: Total de *tweets* por polaridade (criado por este autor).

Infelizmente, a acurácia do modelo, ao classificar os dados negativos, está muito abaixo do esperado. Na figura 19, é possível verificar que o modelo estava considerando muitos *tweets* como neutro, o que pode ter ocorrido pela falta de mais amostras ou pela similaridades dos dados negativos e neutros ou pela má classificação manual dos dados.

No trabalho de Filho (2014), o nível de precisão obtido com o modelo de classificação, sobre a base de teste, foi de 74,80% e as porcentagem de acurácia por polaridade que obteve em seu trabalho estão dispostas na Tabela 6, logo abaixo.

Polaridade	Porcentagem
Positivo	82%
Negativo	84%
Neutro	69%

Tabela 6: Porcentagem de acurácia por polaridade - Filho (criado por este autor).

Ao analisar o nível de precisão do nosso modelo com o de Filho (2014), pode-se verificar que no geral, nosso modelo teve uma melhoria de 2,47% ao classificar corretamente as amostras, porém, seu nível de acurácia das amostras negativas foi superior ao dessa análise. Essa diferença pode ter acontecido pela baixa quantidades de *tweets* negativos presentes na base de dados em questão.

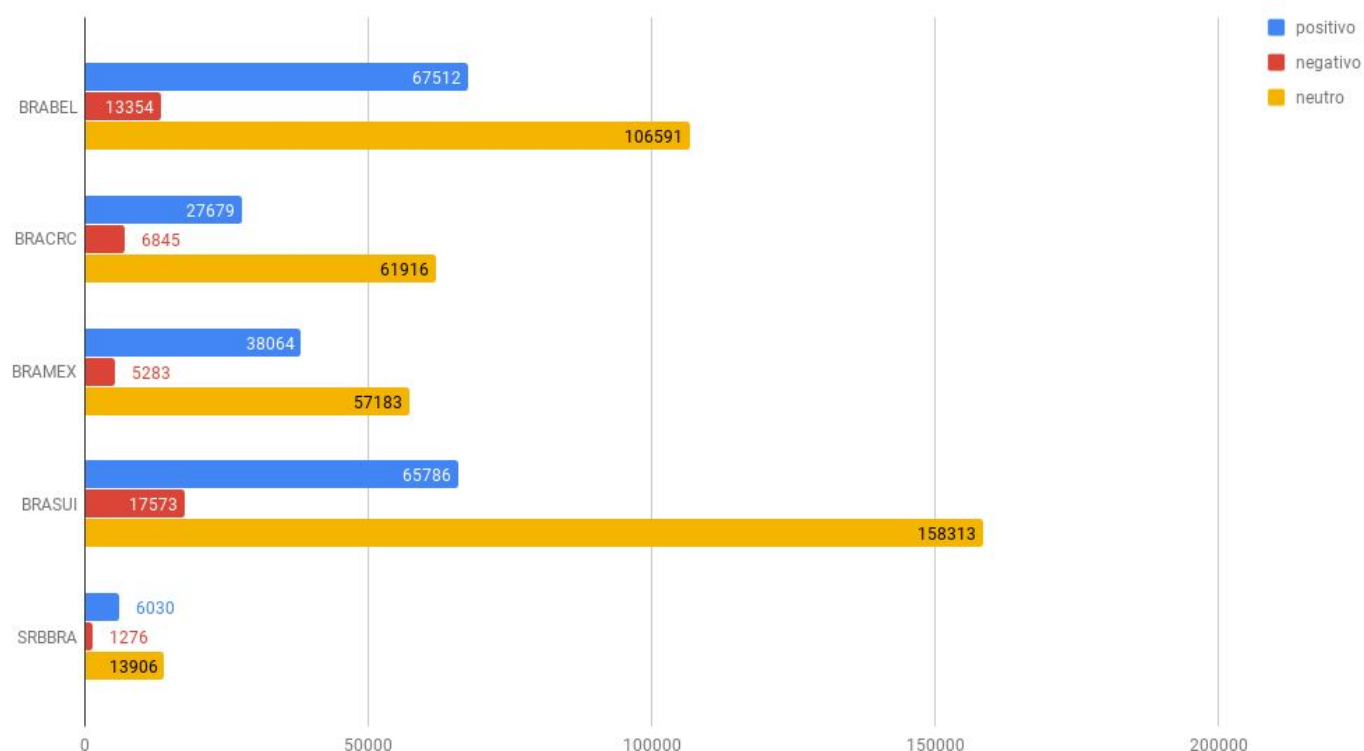
4.4 Análise dos Sentimentos

Após finalizado o modelo de classificação, foram classificados todos os 835.932 *tweets* restantes.

Na Figura 20, é possível observar o resultado obtido ao classificar as amostras restantes. Pode-se verificar que, em todas as partidas, a polaridade que mais se destaca é a neutra, isso se dá pela constante publicação de dados estatísticos dos jogadores e da partida, assim como, propagandas.

Figura 20: Classificação das polaridades por partida.

Classificação por partida



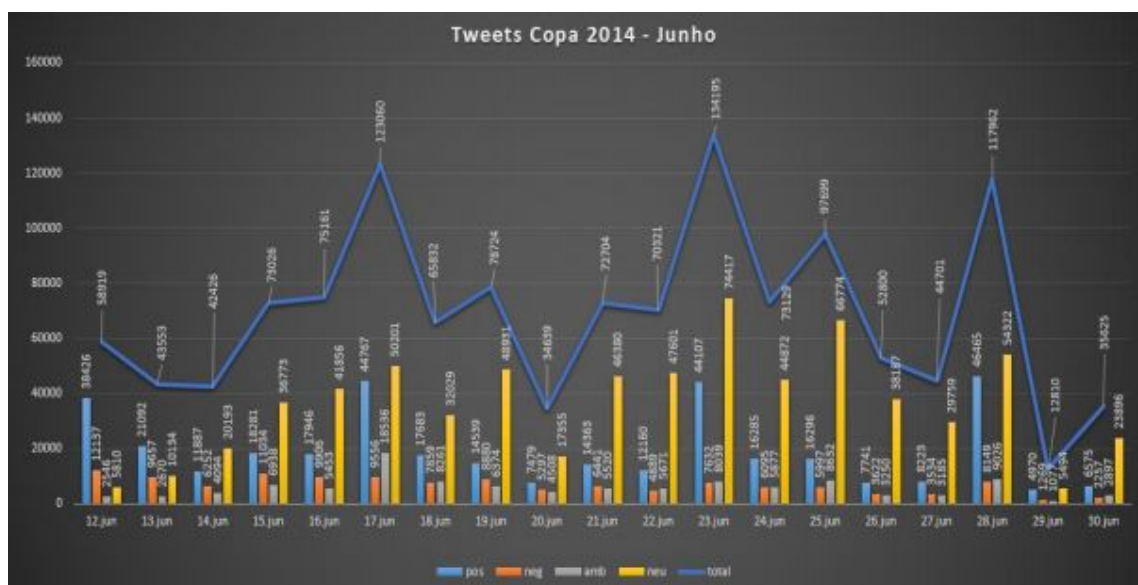
Fonte: criado por este autor.

Se for analisar apenas as polaridades, positivo e neutro, é possível observar que, durante as partidas, os usuários do *Twitter* estavam mais satisfeitos com aquilo que estavam assistindo, do que infelizes com o acontecimento.

Nos resultados da classificação, no trabalho de Filho (2014), dispostos nas Figuras 21 e 22, apresentam uma certa familiaridade com o resultado que obteve-se no presente trabalho. É possível observar que, na grande maioria das partidas, a polaridade que se destaca é o neutro e em segundo o positivo.

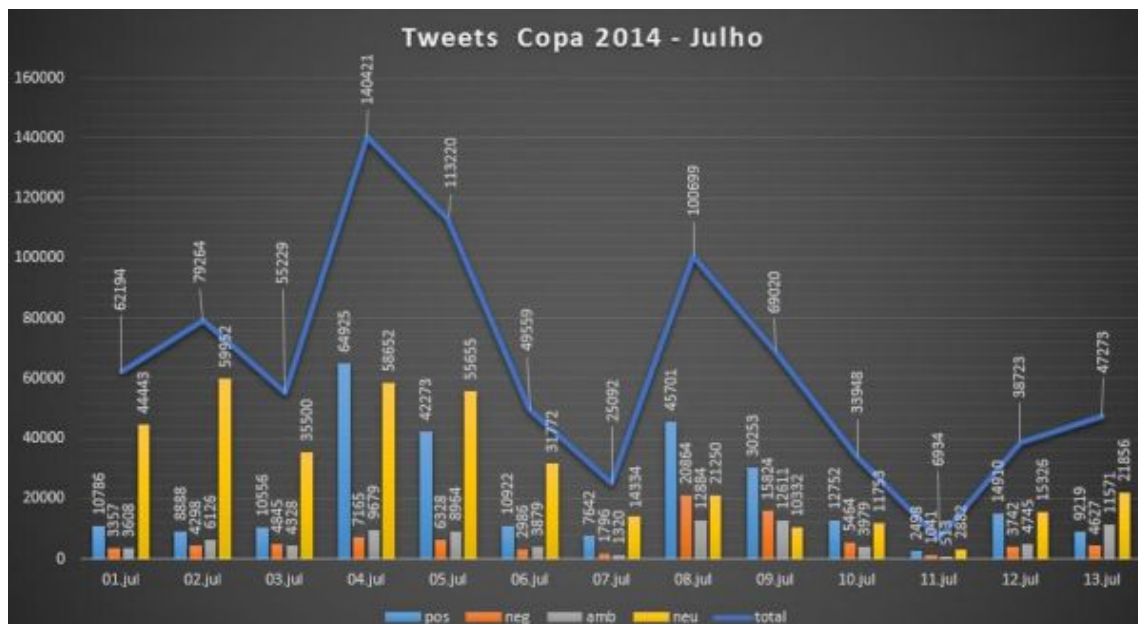
Assim como no presente trabalho, Filho argumenta que essa quantidade de *tweets* classificados como neutro derivam da grande quantidade de dados estatísticos, pertinentes aos jogadores e as partidas. Filho também argumenta que os *retweets* tiveram influência na classificação das amostras como neutras.

Figura 21: Classificação dos *tweets* do mês de junho.



Fonte: (FILHO, 2014).

Figura 22: Classificação dos *tweets* do mês de julho.



Fonte: (FILHO, 2014).

Os dados apresentados neste trabalho e no de Filho (Figuras 20, 21 e 22), mostram que, no geral, a Copa do Mundo de 2014 e as partidas do Brasil na Copa de 2018 foram avaliadas positivamente pelos usuários do *Twitter*.

5. CONCLUSÃO

Neste trabalho, foram utilizados os processos da Mineração de Textos para coletar os *tweets* sobre a Copa do Mundo de 2018, pré-processar esses *tweets* para poder utilizá-los de forma eficiente e criar um modelo de classificação capaz de analisar os sentimentos dos usuários da rede social, *Twitter*. Os dados sobre as partidas do Brasil foram categorizadas entre as três polaridades: positiva, negativa ou neutra.

O modelo de classificação permitiu classificar todas as amostras contidas na base de dados e analisar esses dados para verificar as opiniões dos usuários no decorrer das partidas em que o Brasil participou. Com essa verificação das opiniões, foi possível comparar os resultados obtidos com os dados apresentados por Filho (2014).

Foi interessante verificar que tanto este trabalho quanto o de Filho (2014), apresentaram resultados quase similares, apesar deste trabalho ter conseguido uma acurácia, levemente, superior na classificação das amostras. É possível constatar, também, que a plataforma *Twitter*, consegue mostrar as opiniões dos usuários, mas para este evento em específico, há uma grande quantidade de comentários que não apresentam uma opinião, apenas informações pertinentes a partida.

Uma das conclusões que se pode chegar é de que os métodos utilizados neste trabalho e no trabalho de Filho (2014), podem ser utilizados para analisar os sentimentos dos usuários que utilizam a rede social, *Twitter*, em específico, para este tipo de evento.

REFERÊNCIAS

Allen, James. “Natural Language Understanding (2nd ed.)”. The Benjamin/Cummings Publishing Company. 1995. p.20.

ARAÚJO, Gabriela Denise de; SOUSA, Fernando Sequeira; TEIXEIRA, Fabio; MANCINI, Felipe; DOMENICO, Edvane Birelo Lopes De; GUIMARÃES, Marcelo de Paiva; PISA, Ivan Torres. Análise de sentimentos sobre temas de saúde em mídia social. **Journal of Health Informatics**, v. 4, n. 3, 2012.

BONZANINI, Marco. **Mining Twitter Data with Python**. 2015. Acessado em: 03 de ago. 2018. Disponível em: <<https://marcobonzanini.com/2015/03/02/mining-twitter-data-with-python-part-1/>>.

BRAGA, Daniela Filipa Macedo Moreira da et al. Algoritmos de processamento da linguagem natural para sistemas de conversão texto-fala em português. 2008.

CLOUD, Google. Cloud Translation API Documentation. 2018 Acessado em: 07 de ago. 2018. Disponível em: <<https://cloud.google.com/translate/docs/>>.

FILHO, JOSÉ ADAIL CARVALHO. Mineração de textos: Análise de sentimento utilizando tweets referentes à copa do mundo 2014. 2014.

GOMES, Helder Joaquim Carvalheira. Text Mining: análise de sentimentos na classificação de notícias. **Information Systems and Technologies (CISTI), 2013 8th Iberian Conference on**. Lisboa. 2013.

GONZALEZ, Marco; LIMA, Vera Lúcia Strube. Recuperação de informação e processamento da linguagem natural. In: **XXIII Congresso da Sociedade Brasileira de Computação**. [S.l.: s.n.], 2003. v. 3, p. 347–395.

GOOGLE. Quotas & Limits. 2018. Acessado em: 10 de set. 2018. Disponível em: <<https://cloud.google.com/translate/quotas>> .

HEARST, Marti. What is text mining. **SIMS, UC Berkeley**, 2003.

HOTH, Andreas; NÜRNBERGER, Andreas; PAASS, Gerhard. A brief survey of text mining. In: CITESEER. **Ldv Forum**. [S.l.], 2005. v. 20, n. 1, p. 19–62.

Leung, K. Ming. Naive bayesian classifier. *Polytechnic University Department of Computer Science/Finance and Risk Engineering*. 2007.

LIDDY, Elizabeth D. Natural language processing. 2001.

LIU, Bing. Sentiment analysis and opinion mining. **Synthesis lectures on human language technologies**, Morgan & Claypool Publishers, v. 5, n. 1, p. 1–167, 2012.

MALHEIROS, Yuri; LIMA, George; TINTO-PB-BRASIL, Rio. Uma ferramenta para análise de sentimentos em redes sociais utilizando o senticnet. 2013.

MILLMAN, K Jarrod; AIVAZIS, Michael. Python for scientists and engineers. **Computing in Science & Engineering**, IEEE, v. 13, n. 2, p. 9–12, 2011.

MINER, Gary; IV, John Elder; HILL, Thomas. **Practical text mining and statistical analysis for non-structured text data applications**. [S.l.]: Academic Press, 2012.

MYSQL, AB. **MySQL**. 2001.

PEDREGOSA, Fabian; VAROQUAUX, Gaël; GRAMFORT, Alexandre; MICHEL, Vincent; THIRION, Bertrand; GRISEL, Olivier; BLONDEL, Mathieu; PRETTENHOFER, Peter; WEISS, Ron; DUBOURG, Vincent et al. Scikit-learn: Machine learning in python. **Journal of machine learning research**, v. 12, n. Oct, p. 2825–2830, 2011.

POSTGRESQL. **What is PostgreSQL?** 2018. Acessado em: 28 de jul. 2018. Disponível em: <<https://www.postgresql.org/about/>>.

ROSEN, Aliza. **Giving you more characters to express yourself**. 2017. Acessado em: 24 de mai. 2018. Disponível em: <https://blog.twitter.com/official/en_us/topics-product/2017/Giving-you-more-characters-to-express-yourself.html>.

SANTOS, W. P. Silva. Análise dos Tweets sobre a Black Friday através da Mineração de Texto e Análise de Sentimentos. Rio de Janeiro, RJ - Brasil. 2016.

STATS, Internet Live. **Twitter Statistics**. 2018. Acessado em: 10 de abr. 2018. Disponível em: <<http://www.internetlivestats.com/twitter-statistics/>>.

TWITTER. **Connecting to a streaming endpoint**. 2018. Acessado em: 21 de julh. 2018. Disponível em: <<https://developer.twitter.com/en/docs/tweets/filter-realtime-/guides/connecting.html>>.

Zhang, Harry. **The optimality of naive Bayes**. AA 1, no. 2, p. 3, 2004.