Renata Maciel Camillo

HARRY POTTER E A ANÁLSE DE SENTIMENTOS DO TWITTER

Monografia de Final de Curso

30/07/2020

Monografia apresentada ao Departamento de Engenharia Elétrica da PUC/Rio como parte dos requisitos para a obtenção do título de Especialização em Business Intelligence.

Orientadores:

LEONARDO MENDONZA

EVELYN BATISTA

Agradecimentos

Meus sinceros agradecimentos à minha família, meus filhos, Catarina e Matias, que nasceram no meio dessa empreitada e a minha esposa Antonia, que acompanhou diariamente a correria de pesquisa, trabalho e bebês recém-nascidos. Valeu a pena!

RESUMO

A Análise de Sentimentos, campo da Mineração Textual, que cresce constantemente com os avanços nas áreas de aprendizado de máquinas, processamento de linguagem natural e das redes sociais. O Twitter, maior microblog atual em números de interações, é utilizado por seus usuários como local de busca por avaliações de produtos, serviços, notícias e também de publicações próprias. Com o objetivo de avaliar como se comportam os comentários a respeito do tema, personagem e livro "Harry Potter", este trabalho abordará os conceitos macros de *machine learning* e conceitos que permeiam essa ciência. Como ferramenta, foi utilizado o software RapidMiner e a linguagem de programação Python, capazes de tratar e inferir dados na forma não estruturada.

Palavras-chave: aprendizado de máquinas, RapidMiner, Twitter, análise de sentimento, Python

ABSTRACT

Sentiment Analysis, a field of Textual Mining, is growing constantly with the advances in the areas of machine learning, natural language processing and social networks. Twitter, currently the largest microblog in terms of number of interactions, is used by its users as a place to search for reviews of products, services, news and also for their own posts. With the goal of assessing how the comments behave with regards to theme, character and book “Harry Potter”, this work will deal with the macro concepts of machine learning and the concept that permeate this science. As tools, Python and the software RapidMiner were used, capable of treating and inferring data in unstructured form.

Key- words: machine learning, RapidMiner, Twitter, sentiment analysis, Python

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

1. INTRODUÇÃO

Seguindo os passos de todas as evoluções tecnológicas, sociais, políticas e econômicas dos últimos anos, as redes sociais passaram por diversas transformações e possuem hoje um papel importante nos estudos de diferentes áreas, entre elas a comportamental. Os usuários são induzidos frequentemente a comprar, avaliar e opinar produtos, serviços, entretenimentos, esportes, qualidade de vida, etc., de acordo com seus gostos. Impulsionados a dar opinião acerca de qualquer tema que os convenha, esses usuários adquiriam, por meio do *boom* das redes sociais, *status* de freguês. Segundo levantamento do Kantar, em 2016[1] 46% das exibições da TV apresentaram correlações entre volume de audiência e opiniões apresentadas no Twitter. Já durante o primeiro semestre de 2020[2], devido principalmente à pandemia do novo Corona vírus (Covid-19), o e-commerce teve aumento de 45% por todo Brasil. Segundo outro levantamento[3], diariamente são enviados aproximadamente 500 milhões de *tweets* por dia no mundo, contabilizando também os chamados RT (*retweet)*. Tais exemplos apenas servem para contextualizar o papel que as redes sociais podem ocupar se analisadas de perto pelas instituições e empresas que desejam extrair *insights* e como a marca e/ou produto está sendo compartilhado entre as diversas comunidades.

Alcançar esses objetivos só é possível se conseguirmos extrair da Web os dados necessários para nutrir redes de conhecimentos. Para essa tarefa, utilizamos a linguagem natural, aquela empreendida pela sociedade humana para comunicação, mais precisamente a NLP – processamento de linguagem natural -, para captar as nuances e extrair informação dos dados "jogados" no ambiente virtual. Já a Análise de Sentimentos, visa identificar se essas informações coletadas possuem opinião. Segundo o dicionário Aurélio, a palavra opinião pode ser definida da seguinte forma:

---

[1] Diponível em: <https://br.kantar.com/m%C3%ADdia/%C3%A1udio,-texto,-tv-e-v%C3%ADdeo/2016/abril-coment%C3%A1rios-no-twitter-podem-influenciar-n%C3%BAmeros-de-audi%C3%AAncia/>

[2] Disponível em: <https://br.kantar.com/mercado-e-pol%C3%ADtica/sa%C3%BAde-e-esporte/2020/thermometer-ed9/>

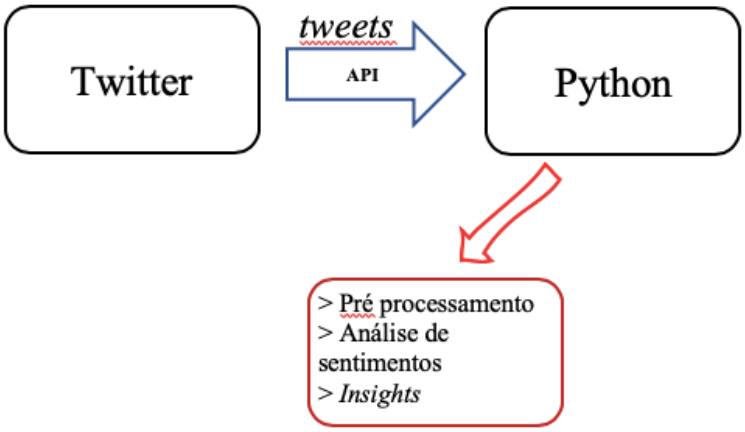
[3] Disponível em: <https://www.dsayce.com/social-media/tweets-day/>

> "Modo de pensar; aquilo que se **pensa em relação a um assunto ou pessoa**; parecer ou ponto de vista: não tenho opinião sobre esse assunto.
> Demonstração de um pensamento pessoal em relação a algo ou alguém; avaliação.
> O que se diz **sem comprovação, fundamento** ou confirmação: sua opinião não comprova os fatos.
> Ponto de vista regulamentado; juízo formado; conceito: não se deve dar uma opinião sem conhecer o assunto.
> **Pensamento comum**; senso habitual de um grupo de pessoas.
> [Informal] Característica de quem não desiste; teimosia: menina de opinião forte!"

Ou seja, a opinião é algo inerente aos seres humanos pois demonstram sentimento em relação a um fato, que não precisa vir acompanhado de comprovação e nem ter uma verdade dita universal. Podem ser formados, através dela, grupos sociais que possuem opinião em comum, ou seja, que enxergam verdades em comum. Apesar de os desdobramentos da opinião em torno de um fato/produto não ser o foco deste trabalho e sim como ela é capturada da Web, é importante frisar o quanto ela pode influenciar não apenas um individuo, mas a sociedade como um todo.

Para que essa rede de análise seja montada, é necessário termos ferramentas da computação que possibilitem a conexão com os repositórios da Web e, além do mais, possuam os algoritmos necessários para que a análise seja feita e daí, ter insights extraídos. Entre elas, utilizaremos as API's (*Application Programming Interface* no original) e o Google Colab, este último para escrever e rodar o código na linguagem de programação Python. A primeira visa possibilitar que aplicativos diferentes "conversem" e troquem informações entre si. No nosso caso, como já mencionado acima, utilizamos as disponibilizadas pelo Twitter para acessar os *tweets* (como são chamados os textos postados na rede) e seus metadados. Segundo Rocha (2020), metadados são os dados sobre o dado – objeto - de forma estruturada. Já com a segunda, utilizaremos o Colab para rodar, via *browser*, o código de programação criado com as etapas de mineração textual que serão descritas ao longo dos próximos capítulos.

Figura 1 - Envio de dados



Os pontos apresentados até aqui serão utilizados, em diferentes níveis, ao logo deste estudo. Porém, antes de mais nada, falaremos um pouco das motivações e objetivos que esperamos alcançar.

1.1 Motivação

Desde que foi lançado em 1997, a saga do personagem Harry Potter trouxe impactos diretos na cultura pop universal. Novos formas de escrever histórias, como *fandons*, *spin-offs* - como *Pottermore*[4] - e produtos de turismo, bonecos, fantasias, parques temáticos, etc., deram ao universo criado por J.K Rowling um *status* diferenciado em relação a outros sucessos literários.

Com as redes sociais, a separação entre vida pessoa do escritor e a história ficcional podem trazer contratempos positivos ou negativos de acordo com o que está circulando de informação pela Internet. Será que as opiniões expostas dos criadores alteram a forma como os produtos são avaliados pelos usuários? O teor dos comentários expressa sentimentos? Será a exposição, no caso Harry Potter, da autora influência nos sentimentos dos usuários com relação ao personagem?

A escolha por utilizar a linguagem Python para extrair, processar e inferir os dados do Twitter, faremos a análise dos *tweets* – atualmente com um limite de 230 caracteres por postagem – que possuem o termo “Harry Potter” em seu corpo.

[4] Disponível em: <https://www.wizardingworld.com/>

As técnicas apresentadas de ETL e processamento de linguagem natural vão auxiliar nossa higienização, inferência e conclusão.

1.2 Objetivo

O objetivo desta pesquisa, conforme já exposto, visa analisar os sentimentos presentes nos *tweets* enviados sobre o personagem/livro/filme Harry Potter. A divergência de sentimentos do publico consumidor em relação à sua criadora pode estar ou não impactando o sentimento em relação ao personagem.

Planejamos desenvolver

- Seleção aleatória de *tweets* recentes, populares e enviados na língua inglesa
- Análise de um número considerado aceitável – 4,5 mil - de *tweets* que possibilite uma melhor certeza da polaridade do sentimento com as bibliotecas disponíveis
- Explicar conceitos iniciais de aprendizado de máquinas, possibilitando descobertas de melhores ferramentas e entendimento do estudo realizado
- Contextualizar o papel das redes sociais para os usuários e produtores de conteúdo, como sites, lojas, jornais, formadores de opinião, etc.
- Verificar se o termo “Harry Potter” está sendo comentado de forma positiva, negativa ou neutra no Twitter

1.3 Descrição do trabalho

O presente trabalho está dividido da seguinte forma: no capítulo 2 veremos conceitos introdutórios de aprendizado de máquinas, divido em aprendizado não supervisionado e supervisionado. Em seguida falaremos de como é organizado, em etapas, o projeto de mineração textual (coleta, pré-processamento, indexação, mineração, análise). No capítulo 4, explicaremos do que se trata NLP (processamento de linguagem natural) e 5 como será feita a o foi coleta dos dados.

No capítulo 6 falaremos do processo de pré-processamento, que mesmo sendo uma etapa do projeto de mineração, pela importância e práticas exclusivas do

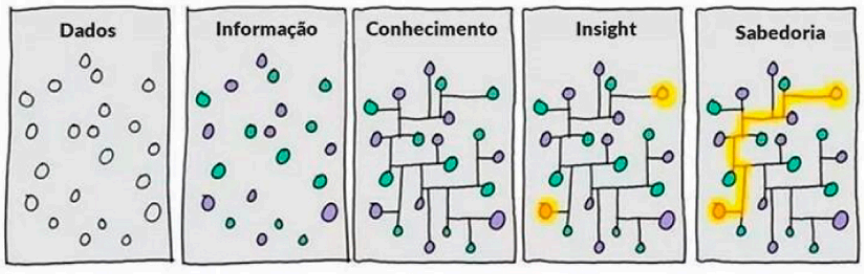
tratamento textual, terá um capítulo exclusivo. Nos capítulos 7, falaremos da análise de sentimentos propriamente dito. Nos capítulos finais mostraremos o projeto e o passo a passo do script em Python.

2. APRENDIZADO DE MÁQUINAS

> "A característica do aprendizado pode dotar aos agentes de capacidade de desenvolver sua autonomia, pois com o aprendizado o agente está sempre apto a se adaptar aos ambientes." (Lopez, pág. 21)

Aprendizado é o ato de aprender um ofício, uma arte ou ciência (dicionário Michaelis online). Já **conhecimento** é o ato de conhecer por meio da razão e/ou da experiência (dicionário Michaelis, online). No aprendizado de máquinas, cientistas e pesquisadores buscam na natureza formas de otimizar processos de modo computacional para que possamos extrair conhecimento, até mesmo sabedoria, de dados que antes não existiam ou que não tínhamos consciência de existir.

Figura 2 - Evolução no processamento de dados



Fonte: Fumsoft

Existem 3 elementos citados por Mitchel (1997) que possibilitam os humanos a ensinarem para máquinas, o que possibilitaria a elas o aprendizado: experiência, tarefa e medida de desempenho. Para Barreira (2013, pág. 16, grifo nosso):

> "Ao trabalhar com a questão da aprendizagem na área de computação, muitas teorias têm como base a utilização do raciocínio indutivo. Esse tipo de aprendizado indutivo pode ser assim definido (MITCHEL, 1997, p.2, tradução nossa): "um programa de computador aprende a partir de uma **experiência** E em relação a algumas classes de **tarefa** T e medidas de **desempenho** P, se esse desempenho na tarefa T, conforme medido por P, melhora a experiência E".

Ou seja, podemos observar o "mundo real", dali tirar experiências e, através da indução, apresentar uma tarefa ao computador. Esse processo de ensinar, será divido aqui em duas formas: o aprendizado supervisionado e o não supervisionado. O primeiro é utilizado quando o conjunto de dados apresentado ao computador

possui rótulos, ou seja, classes, previamente identificados. Já no segundo, o conjunto não possui essa característica e o computador precisa identificar semelhanças entre os dados para instruir-se.

Para cada tipo de aprendizado, temos empregabilidades diferentes, podendo construir inúmeras ferramentas e extrair infinitos *insights*.

2.1 Aprendizado não supervisionado

O método não supervisionado é utilizado para problemas em que temos um conjunto de dados sem rótulos – classes - como por Associação[5] e por Agrupamento, conforme mostrado na figura 4 – *clusters*. Não utilizaremos dele neste estudo, mas não há dúvidas de que é um divisor de águas no que diz respeito ao que hoje entendemos como inteligência artificial. Uma das qualidades é a não necessidade dos especialistas destinados a categorizar e rotular os dados previamente, função essa do próprio aprendizado.

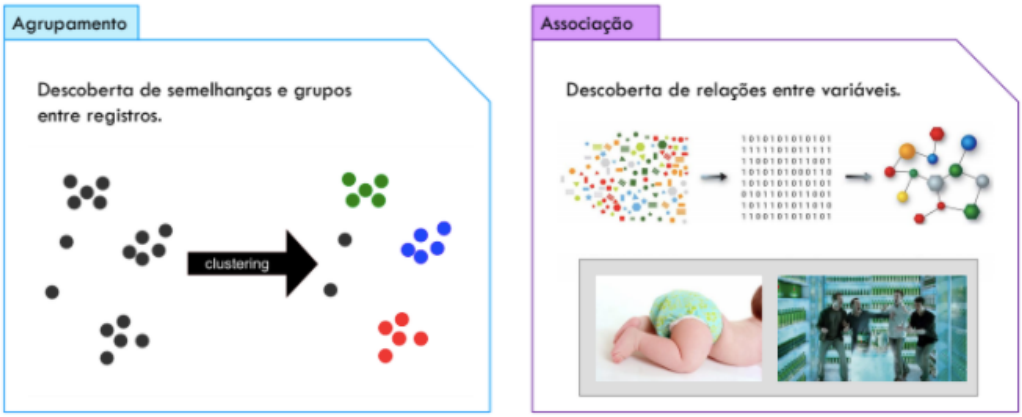
A *clusterização* – agrupar de acordo com as similaridades dos dados – é vista como uma das qualidades deste modelo, pois, além de ser menos trabalhoso no que diz respeito aos grupos de dados, funciona de acordo com necessidades intrínsecas aos próprios dados. A não supervisão acaba trazendo a economia de energia desprendida no processo de adequação dos requisitos e contribui para que os dados utilizados sejam de fato aqueles necessários, diminuindo também a redundância. Como pré-requisito para esse processo, segundo Filho (2017, pág. 25, **apud MILLIGAN, 1996**) **(HENNIG; LIAO, 2013) (LUXBURG; WILLIAMSON; GUYON, 2012)** temos que:

> "[...] escolher os objetos aos quais se deseja agrupar, escolher os valores a serem utilizados, padronizar esses valores, escolher a medida de similaridade, escolher o método de *clustering*, decidir o número de clusters (alguns métodos exigem o número *clusters* como parâmetro de entrada) e finalmente testar, interpretar e validar os *clusters*.[...] A tarefa de selecionar um método e interpretar seu resultado implica no bom entendimento dos dados para que um método apropriado possa ser usado"

---

[5] Disponível em: <https://exame.com/revista-exame/o-que-cerveja-tem-a-ver-com-fraldas-m0053931/ > para saber mais a respeito do estudo realizado sobre vendas de fraldas e cerveja em um grande mercado americano.

Figura 3 - Aprendizado não supervisionado



Fonte: KHOELER, 2018

2.2 Aprendizado supervisionado

Podemos utilizar o método supervisionado para problemas de classificação, regressão e previsão de séries temporais, conforme apontado na figura 5. Na classificação, basicamente criamos categorias discretas, como masculino e feminino, frutas e verduras, 0 ou 1, etc. Na regressão, utiliza-se rótulo contínuo e com séries temporais, o tempo entra na equação.

Figura 4 - Aprendizado supervisionado

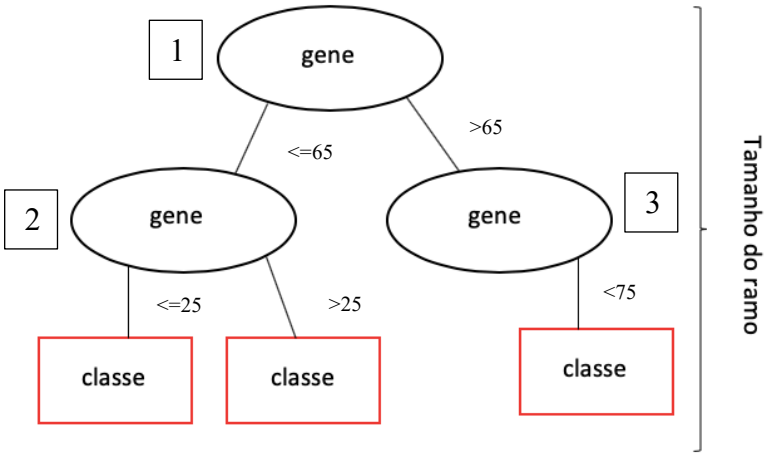
Fonte: KHOELER, 2018

Para Lopez (2010, pág. 21), o aprendizado tem por objetivo tornar a ação executada e desejada o mais próximas possíveis para que seja, em seguida, avaliado o nível do erro. Assim, aplica-se peso de ajuste até o erro mínimo aceitável seja encontrado. Primeiro, define-se o conjunto de dados que será o saber, com entradas

e saídas entendidas a priori. Em seguida, a saída desejada (treinamento) é entregue ao algoritmo, que será ajustado de acordo com os erros apresentados. O erro, que é definido na combinação da resposta obtida e deseja, é aqui o definidor dos pesos que serão validados.

Para mineração de textos – *text mining* – podemos resolver esse tipo de problema utilizando o método de classificação, pois podemos categorizar os sentimentos extraídos de acordo com a polaridade das palavras, como positivo, negativo e neutro – classificação esta comumente utilizada -. A Classificação nada mais é que rotular os dados de acordo com uma característica específica. Por exemplo, na figura abaixo podemos avaliar um grupo de pessoas que está inserido no mercado de trabalho. Do gene 1 ao gene 2, selecionamos as pessoas que possuem até 65 anos, ou seja, que ainda não possuem idade média para aposentadoria (a diferença entre homens e mulheres não vem ao caso). No gene 2, separamos aqueles que possuem até 25 anos e que estão, em média, recém-chegados no mercado de trabalho, e aqueles que são maiores que 25 e, aparentemente, alguma experiência. No gene 3, isolamos aqueles que possuem mais de 65 e menos de 75, já em idade de aposentar, mas que por algum motivo ainda estão ativos. Já as classes são os nomes que daremos a esses grupos de dados. No caso, podemos classificar como na tabela 4.

Figura 5 – Exemplo de classificação



Os genes e, consequentemente as classes, são geradas a partir das regras inseridas de acordo com o corte nos cortes que estamos pesquisando. Como exemplo, selecionamos um grupo de dados e para corte. As classes poderiam ser rotuladas de grau 1, grau 2 e grau 3 para analisar o mercado de trabalho.

Tabela 1 – Exemplo de classificação

Corte	Rótulo - Classe
Menores de 25 anos	Adulto jovem
Maiores de 25 anos e menores que 65 anos	Adulto maduro
Maiores que 65 anos e menores que 75 anos	Aposentados Ativos

3. PROJETO DE MINERAÇÃO TEXTUAL

De acordo com Aranha (2007), podemos dividir projetos de mineração textual em 5 processos.

Figura 6 - Projeto de mineração textual

Fonte: ARANHA, 2007

a) **Coleta**: processo no qual há a coleta de dados oriundos das bases de dados selecionadas para utilização.

b) **Pré-processamento**: de acordo com Brito (2017, pág. 27), o principal objetivo de pré-processar um texto, consiste na filtragem e limpeza dos dados, eliminando redundâncias e informações desnecessárias para o conhecimento que se deseja extrair (*apud* Gonçalves et al., 2006).

c) **Indexação**: utilizado para otimizar a busca e recuperação de termos, criando palavra-chave e filtros

d) **Mineração**: Processo de *machine learning* propriamente dito, onde os algoritmos de inferência atuam no texto, extraindo informações.

e) **Análise**: Nesse momento vemos a interpretação dos dados que foram colhidos, tratados e minerados. Para Brito (2017, pág. 27) esse é o ponto onde ocorre a avaliação do processo como um todo.

Apesar de esses processos ocorrerem, levando em consideração as particularidades do que está sendo tratado, em outros tipos de projeto de mineração, no caso de textos extraídos de ambientes virtuais precisamos das API's para que a

coleta seja viável. Na etapa de pré-processamento realizamos as atividades de tokenização, *stemming* e *stopwords* para higienizar e padronizar ao máximo as palavras, aumentando as chances de melhores resultados na etapa de Mineração e Análise dos resultados. A indexação, apesar de transparente no código e não ser o nosso foco, é vital para a recuperação das informações e serve como ponte para o processo de mineração.

3.1 Processamento de Linguagem Natural

A NLP, processamento de linguagem natural, é o campo da ciência computacional que busca entender as relações linguísticas humanas para que as mesmas sejam utilizadas pela inteligência artificial (AI). Através do aprendizado de máquinas, o computador adquire a capacidade de ler, entender e inferir sobre o que foi escrito em um documento qualquer.

Levando em conta o grande crescimento das exigências computacionais para acompanhar os esforços atuais, como já dito a respeito dos avanços das redes sociais e do aprendizado de máquinas, constatou-se que existe um vácuo de conhecimento máquina/língua x humano/língua. Apesar de enxergarmos esse vazio como natural, visto que a inteligência computacional é baseada naquilo que nós mesmos apresentamos, é preciso que a apresentação das regras linguísticas e do máximo de dados possível seja feito para que as falhas e a não identificação de termos e expressões sejam minimizados ao máximo. Para Neto, Tonin e Prietch (2010), problemas como os relatados acima só serão devidamente tratados ao serem desempenhados 3 tipos de análise: **morfológico**, quando estudamos a forma da língua, ou seja, como ela está sendo estruturada, **sintático**, quando verificamos a relação da unidade palavra em relação a outras palavras, e **semântico**, análise do sentido e interpretação das sentenças. Também segundo Neto, Tonin e Prietch (2010, pág. 4):

> *"*Se, em relação ao tratamento do léxico, os dicionários utilizados pelos sistemas de processamento não forem adequados, quer do ponto de vista da sua cobertura *l*exical, quer, do ponto de vista da formalização e sistematização da informação lingüística, isso afetará não só a análise lexical de um determinado texto, mas *t*ambém todas as fases de processamento subseqüentes*."*

Sendo assim, com essas três categorias de análise, podemos atuar em cada uma de forma diferente. No primeiro caso, morfológico, utiliza-se separadores,

como espaços e pontuações, para definir a palavra, expressão, sentença, frase. No segundo caso, sintático, utilizamos da classificação para dar nomes a classes de palavras, agrupando-as de acordo com as regras gramaticais. No terceiro caso, o nível semântico avalia o significado das palavras e dos agrupamentos - expressões – formados.

Ainda segundo Neto, Tonin e Prietch (2010, pág. 5), os progressos da NLP permitem ainda uma 4ª análise, chamada pragmática. Nela são analisadas o todo e não apenas as partes descritas até aqui. Para eles, as palavras podem se associar através de dois tipos de relações: paradigmáticas (significado) e sintagmáticas (quantidade de vezes que a palavra é localizada no discurso). Ou seja, as paradigmáticas analisam o significado enquanto a sintagmáticas verificam o contexto na qual a mesma está empregada.

Os levantamentos das regras devem ser feitos de acordo com o idioma utilizado ao longo da análise pois, como verificamos até aqui, cada idioma possui particularidades e regras que não são aplicadas iguais nas diferentes línguas. Para o idioma inglês, devido a investimentos aplicados na área desde a década de 1950, é possível obtermos uma melhor resposta computacional.

Já a identificação de entidades (REM), apesar de banal para nós seres humanos, possui dificuldades como o duplo sentido de significados. Enquanto "Palmeira" pode ser o nome de uma espécie de árvore, pode também ser um sobrenome humano, por exemplo. No inglês, termos como "*cloud*" podem ter relação com o tempo/clima, mas também com centros de custódia de documentos ou quando falamos da oferta do serviço de custódia oferecido. Logo, conforme Aranha (2007, pág.71) aponta, é de grande importância a especificação, além de nomear o objeto do mundo real de trabalho. Grande parte da informação de uma nova notícia é proveniente de novos nomes, ou relacionamentos entre novas combinações de nomes.

Para Aranha (2007, pág. 60), apesar de o principal objetivo ser o reconhecimento e classificação de entidades (nomes de pessoas, lugares, estabelecimentos, instituições, etc.), para que isso seja bem aproveitado é necessário que outras atividades em conjunto sejam realizadas com a NLP. Durante a etapa de pré-processamento, é fundamental que seja realizado a tokenização (tópico 6.1) e lematização/*stemming* (tópico 6.3).

3.2 Coleta de dados

Essa é a etapa que busca observar e colher os registros que serão utilizados ao longo da análise do objeto de estudo. Ela pode ser feita de diferentes formas e abordagens, com diversos tipos de dado – estruturado ou não -, documento – web, papel, log de sistema, etc. -, coleção – conjunto de dados específicos -. Porém, antes de iniciarmos essa etapa, é necessário definir como será feito e quais atributos serão carregados. Quando buscamos, por exemplo, dados oriundos da internet, precisamos verificar se o formato do dado será satisfatório na integração com outras ferramentas. Ou seja, precisamos nos certificar que será possível ler os dados gerados.

Através das API's desenvolvidas pelo próprio Twitter, conseguimos extrair as informações disponíveis nos *tweets*. Aqui, utilizaremos um *script*, também em linguagem Python, desenvolvido pelo Laboratório de Inteligência Computacional Aplicada da PUC-Rio. Este irá acessar a API do microblog e exportar os dados de forma estruturada e em formato **.csv**[6]. Pesquisaremos o personagem "Harry Potter", em lingua inglesa (en) e oriundos dos Estados Unidos (US)

Por uma limitação de bibliotecas disponíveis para análise em língua portuguesa, faremos a coleta dos *tweets* em língua inglesa, possibilitando assim a inferência dos dados nas próximas etapas do projeto. Apesar de analisarmos os textos enviados pelos usuários, temos disponíveis também informações de localização, data, quantidade de *retweet* – reutilização do *tweet* enviado por outro usuário, pode ser usado para reiterar ou discordar de uma opinião[7] -.

---

[6] *Comma-separated values*, as colunas são separadas por virgulas

[7] As análises de popularidade de um *tweet* específico, por exemplo, deve levar essa informação em consideração pois essa quantidade pode representar um momento "viralizado" ou "cancelado", popular positivo ou popular negativo, respectivamente.

Figura 7 – Forma estruturada dos dados coletados do Twitter

```
import pandas as pd
tweets = pd.read_csv('tweets.csv', engine='python')
tweets.head()
```

	Unnamed: 0	date	tweet	username	retweet	nlikes	nreplies	nretweets	near
0	0	2020-05-31 23:44:06	Yeah my hubby made it through the first 3 Harr...	Felinefemale	False	1	1	0	United States
1	1	2020-05-31 23:38:43	Harry Potter movies?	Felinefemale	False	0	2	0	United States
2	2	2020-05-31 23:22:49	No thanks, I am watching a Harry Potter rerun ...	mrboneheaddave	False	0	0	0	United States
3	3	2020-05-31 22:54:15	The top 4 books on Amazon are Harry Potter. \n...	NathanWurtzel	False	0	1	0	United States
4	4	2020-05-31 22:49:47	During our new daily exercise routine, inspire...	CristaGalli22	False	0	0	0	United States

Tabela 2 – Rótulos dos dados coletados

Date	Tweet	Username	Retweet	Nlikes	Nreplies	Nretweets	Near
Data e hora do nvio do *tweet*	Texto	Usuário	TRUE – é um *retweet* FALSE – não é um retweet	Número de *likes*[8]	Número de respostas dada ao *tweet*	Número de vezes que foi replicado	Localização

3.3 Pré-processamento

Essa etapa é essencial para todo projeto de *data mining*, porém, ao lidarmos com texto, a importância de ser feito conscientemente é vital para o resto do projeto. A língua escrita e falada é, independente do idioma, algo vivo – orgânico - e está em constante mutação. Quando lidamos com redes sociais, precisamos nos adequar ao fato de que não existem padrões linguísticos e mais, gírias e abreviações, além de circularem livremente, são inovadas a todo momento. Utilizaremos as seguintes bibliotecas: NLTK, *TokTokTokenizer* (NLTK), *Spacy*, *BeautifulSoup* (Bs4), Re, *Contractions*, *Unicodedata*.

Faremos usos de técnicas de processamento de texto que são essenciais para que os dados sejam, ao máximo, higienizados e padronizados para que o algoritmo consiga extrair informações. Segundo Junior (2007, pág. 30):

> "Pré-processamento é a etapa realizada imediatamente após a Coleta, com o objetivo de se obter alguma estrutura para a massa textual. Pré-processar textos é, por muitas vezes, o processo mais oneroso da metodologia de MT, uma vez que não existe uma única técnica que possa ser aplicada para a obtenção de uma representação satisfatória em todos os domínios."

[8] Também analisa popularidade, indica concordância com o texto postado

Figura 8 – *Wordcloud* com 300 principais termos

3.3.1 Tokenização e *bag of words*

É a etapa onde as palavras são reduzidas ao mínimo sem que tenha perda de sentido no texto. De acordo com a estrutura da frase, pode ser representado pela própria palavra ou por delimitadores como pontuação, espaço e tabulação do arquivo. Para os conteúdos oriundos das Web, em especial das redes sociais, não raro, pontuações podem ser utilizadas para representarem estados de espírito. Nesses casos, certas ações podem ser feitas. A remoção, por exemplo, de caracteres especiais pode auxiliar o software ao longo da análise e ganhar conhecimento a respeito da estrutura de linguagem, removendo ou substituindo símbolos por texto. Em outros casos, essa remoção pode justamente descaracterizar representações de emoção, como por exemplo o *emoji* ':('. Exemplos desse tipo podem acabar sendo removidos partes ou a totalidade do mesmo.

Segundo Soares (2009, pág. 17), para uma extração eficiente de *tokens*, diversos aspectos da linguagem devem ser considerados, inclusive a diferença entre letras maiúsculas e minúsculas, pois geralmente indica se a palavra está no meio ou no início da frase. Contribui também para o processo de *bag of words*, onde cada

palavra é transformada em um atributo, que recebe valor baseado em sua frequência. Também segundo Soares (2009, pág. 15), essa abordagem pode ser utilizada na extração da Web, pois transforma dados não estruturados em estruturados, mesmo que ignore o contexto na qual é apresentada e utilizada.

Com a linguagem Python, podemos buscar bibliotecas e modelos pré treinados para idiomas específicos em repositórios como o GitHub e o Kaggle. Infelizmente, modelos em português não são, por enquanto, acessíveis. Projetos específicos ou de agentes que não disponibilizam facilmente seus dados, acabam criando bases próprias, únicas e para fins específicos. De qualquer forma, quando lidamos com o idioma inglês, por exemplo, onde existem diversas fontes e catálogos de diversos repositórios, conseguimos, com poucas linhas, acessar esses ambientes e trabalhar nossos dados. Como já mencionamos acima, faremos uso do TokTokTokenizer, da biblioteca NLTK.

```
“# Carregar tokenizador
nlp = spacy.load('en')
tokenizer = ToktokTokenizer()”
```

3.3.2 Stop Words

As *stop words* são palavras que não possuem valor no contexto da frase e podem ser retirados sem que o sentido do que está sendo dito seja alterado. Artigos e verbos de ligação são exemplos. Na tabele abaixo, identificamos as palavras retiradas, em inglês, dos *tweets* e que, a priori, não alteram a polarização dos sentimentos. Notemos que a palavra sinalizada em vermelho (‘pic’) está por último pois foi adicionada manualmente por nós ao código. Ou seja, podemos utilizar uma lista, aqui a da biblioteca do NLTK, e também adicionar de acordo com a necessidade.

Mantivemos as aspas na tabela 3 por serem *strings*. Para adicionar palavras à lista utilizamos o “**stopword_list.append('PALAVRA')**”. Chamamos a função abaixo para rodar a lista.

```
“def remove_stopwords(text):
  tokens = tokenizer.tokenize(text)
  tokens = [token.strip() for token in tokens]
  filtered_tokens = [token for token in tokens if token.lower() not
in stopword_list]
  filtered_text = ' '.join(filtered_tokens)
  return filtered_text”
```

Ao longo da construção, podemos excluir palavrões, palavras obscenas, com duplo sentido e sinônimos, mas isso deve variar. Por exemplo, para Nielsen (2011, tradução e grifo nossa) foi importante manter "surpresa", que pode ser interpretado com mais de um sentido.

> "A lista de palavras começou com uma série de **palavras obscenas e algumas palavras positivas**. Ela foi gradualmente expandida examinando postagens do Twitter coletadas para COP15, especificamente postagens com alta pontuação de sentimento usando a lista conforme ela crescia. Eu incluí palavras do domínio publico *Original Balanced Affective Word List* do *Greg Siegle*. Depois eu adicionei gírias da Internet olhando no *Urban Dictionary*, incluindo siglas como WTF, LOL e ROFL. As inclusões mais recentes vêm da grande lista de palavras do Steven J. De Rose, *The Compass DeRose Guide to Emotion Words*. As palavras de DeRose são categorizadas, mas não pontuadas de acordo com valor, com valores numéricos. Junto com as palavras de DeRose eu busquei no *Wiktionary* e usei os sinônimos sugeridos para aprofundar mais a lista. Em alguns casos usei o Twitter para determinar em quais contextos a palavra aparecia. Também usei o Microsoft Web *N-gram similarity Web Service* ("agrupamento de palavras com base em similaridade de contexto") para descobrir palavras relevantes. Mas não faço distinção entre categorias de palavras, portanto, para evitar ambiguidades, excluo palavras como *patient, firm, mean, power e frank*. Palavras como "**surpresa**" – com alta "excitação", mas sentimento variável – não foram incluídas na lista de palavras."

Tabela 3 – Lista de *stopwords*

Lista de Stopwords						
'i',	'himself',	'which',	'do',	'or',	'about',	'no',
'me',	'she',	'who',	'does',	'because',	'against',	'nor',
'my',	"she's",	'whom',	'did',	'as',	'between',	'not',
'myself',	'her',	'this',	'doing',	'until',	'into',	'only',
'we',	'hers',	'that',	'a',	'while',	'through',	'own',
'our',	'herself',	"that'll",	'an',	'of',	'during',	'same',
'ours',	'it',	'these',	'the',	'at',	'before',	'so',
'ourselves',	"it's",	'those',	'and',	'by',	'after',	'than',
'you',	'its',	'am',	'but',	'for',	'above',	'too',
"you're",	'itself',	'is',	'if',	'with',	'below',	'very',
"you've",	'it',	'are',	'to',	'again',	'all',	'couldn',
"you'll",	"it's",	'was',	'from',	'further',	'any',	"couldn't",
"you'd",	'its',	'were',	'up',	'then',	'both',	'didn',
'your',	'itself',	'be',	'down',	'once',	'each',	"didn't",
'yours',	'they',	'been',	'in',	'here',	'few',	'doesn',
'yourself',	'them',	'being',	'out',	'there',	'more',	"doesn't",

'yourselves',	'their',	'have',	'on',	'when',	'most',	'hadn',
'he',	'theirs',	'has',	'off',	'where',	'other',	"hadn't",
'him',	'themselves',	'had',	'over',	'why',	'some',	'hasn',
'his',	'what',	'having',	'under',	'how',	'such',	"hasn't",
"won't",	'wouldn',	"wouldn't",	'pic'	'will',	'o',	'y',
's',	'd',	't',	'll',	'can',	'm',	"don't",
'don',	've',	'should',	'now',	"aren't",	'ain',	"isn't",
'just',	're',	"should've",	'haven',	"needn't",	'aren',	'ma',
"haven't",	'shan',	"mightn't",	"mustn't",	"wasn't",	"shouldn't",	'mightn',
'isn',	"shan't",	'mustn',	'needn',	'weren',	'wasn',	'shouldn',
"weren't",	'won',					

3.3.3 Stemming e lemmatize

Já o método *stemming* é utilizando para transformar a palavra ao máximo até que ela se torne apenas o seu radical. Em outras palavras, as flexões são retiradas para que tenhamos o mínimo da palavra. Esse processo, segundo Soares (2009, pág. 17) pode ser entendido como a normalização textual.

Um dos algoritmos de *stemming* mais utilizados é o de Porter. Segundo Alvares (2014, pág. 9), ele é dividido em 5 fases, na qual são aplicadas regras que atuam a remoção dos sufixos mais comuns. Atuando junto com as *stopwords*, o *stemmer* evitaria atuar em palavras pequenas e que fazem parte da lista de remoção. Caso trabalhássemos com um projeto em língua portuguesa, o mais indicado seria o algoritmo *Snowball*[9], que já possui o dicionário para o idioma.

```
“def stemmer(text):
  ps = nltk.porter.PorterStemmer()
  text = ' '.join([ps.stem(word) for word in text.split()])
  return text”
```

Agindo com o mesmo propósito de redução, a lematização também agrupa os termos que são variados de um único lema. Os termos são transformados até atingirem o infinitivo e adjetivos e substantivos em masculino singular - devido as regras de generalização dos idiomas -. Porém, diferentemente do *stemming*, a lematização garante que a palavra reduzida exista no vocabulário do idioma, tendo assim um limite até onde a redução pode ocorrer. Por conta disso, vemos mais

[9] Disponível em: <http://snowball.tartarus.org/algorithms/portuguese/stemmer.html>

exigência computacional e, tendo em mente que nosso projeto está sendo rodado em um ambiente com baixo processamento (em comparação a laboratórios especializados e utilizados apenas para essa finalidade), optamos por apostar no método *stemming*.

Ainda podemos realizar outros tipos de pré-processamento que vão de acordo com as especificidades de cada estudo, como tratamento de acentos - que apesar de serem muito comuns na língua portuguesa, não o são na língua inglesa – adição de termos não cadastrados em listas de *stopwords* (como fizemos acima), retirar contrações de palavras, excesso de espaços e de linhas em branco, etc. Na tabela 3, mostramos um exemplo de tratamento.

```
“ # remove acentuação
 doc = remove_accent(doc)
 # expandir contrações
 doc = expand_contractions(doc)
 # coloca tudo em caixa baixa
 doc = doc.lower()
 # remove linhas em branco
 doc = re.sub(r'[\r|\n|\r\n]+', ' ', doc)
 # remove caracteres especiais
 doc = remove_special_characters(doc)
 # remove linhas em branco
 doc = re.sub(' +', ' ', doc)”
```

Tabela 4 – Antes e depois do pré-processamento

Antes	Depois
Welcome to the Harry Potter At Home hub where you’ll find all the latest magical treats to keep you occupied - including special contributions from Bloomsbury and Scholastic, nifty magical craft videos (teach your friends how to draw a Niffler!), fun articles, quizzes, puzzles and plenty more for first-time readers, as well as those already familiar with the wizarding world. We’re casting a Banishing Charm on boredom![10]	welcome harry potter home hub find late magical treat keep occupy include special contribution bloomsbury scholastic nifty magical craft video teach friend draw niffler fun article quiz puzzle plenty first time reader well already familiar wizarding world cast banish charm boredom

[10] Texto retirado e traduzido do site https://www.wizardingworld.com/collections/harry-potter-at-home.

4. MINERAÇÃO DE DADOS

4.1 Análise de sentimentos

A Análise de Sentimento, ou Mineração de Opinião, é o estudo das emoções, sentimentos, opiniões, presentes em textos independente do seu formato e estrutura. Para Silva (2016, pág. 4, **apud LIU & ZANG, 2012**) "uma opinião é uma declaração subjetiva, com uma visão pessoal que expressa atitude, emoção ou apreciação sobre uma entidade ou um aspecto de uma entidade, um parecer".

Vai utilizar técnicas da mineração de dados, linguística, aprendizado de máquinas, processamento de linguagem natural, para coletar e extrair polaridade (positiva, negativa, neutra) de determinado tópico. Para Brito (2017, pág. 17), a polaridade pode ser extraída através da classificação mesmo que a frase não contenha explicitamente um sentimento. Essa prática permite que sejam quantificadas, qualificadas e avaliadas opiniões publicas acerca de produtos, notícias, tópicos, auxiliando tomadas de decisões por entidades.

Segundo Silva (2016, pág. 4) e Brito (2017, pág. 18), Liu (2012) divide a análise em categorias distintas, pois pode haver diferentes níveis de granularidade:

a) **Do documento**: verificar a totalidade do documento, podendo ele ser positivo, negativo ou neutro.

b) **Baseada em aspectos**: mais específica, verifica cada sentença do documento, relaciona entidade com sentimento

Enquanto o primeiro, do documento, foca no produto que está sendo opinado e ter uma base positiva ou negativa é importante para se ter uma visão geral de como o produto está sendo aceito pela sociedade nele incluindo. Já o segundo, baseada em aspectos, pode ser visto como uma análise mais profunda do mesmo produto, mas focando nas melhorias que podem ser feitas para aumentar a popularidade. Vejamos o exemplo: "Não gostei do filme, mas valeu a experiência de ver o novo formato de vídeo". Apesar de o usuário não ter gostado do filme em si, valeu a pena ter visto. Ou seja, ao enxergarmos apenas a entidade "filme", diríamos ser um comentário negativo. Porém, na segunda parte, a experiência aparece como positiva. Para Aranha (2007, pág. 70), uma das dificuldades é a extração de entidade e dificilmente conseguimos extrair 100% delas presentes em um texto. Sem ela, frases com dupla polarização podem ser um problema.

Os aspectos linguísticos que permeiam essa análise são sensíveis e não podemos esperar que em nível computacional consigamos extrair todas as relações entre as palavras, quiçá o sentido textual de frases postadas em um ambiente descontraído e coloquial, com gírias e expressões particulares de grupos específicos culturais. Esse tipo de situação é uma dificuldade quando treinamos os modelos de aprendizado, pois a constante mutação dos textos impede que um padrão seja mantido por muito tempo.

Apesar dessas dificuldades, o Twitter é uma fonte constante de usuários e possíveis clientes, fazendo com que o interesse nesse tipo de abordagem não se interrompa. Segundo levantamento, 84% dos consumidores que buscam serviços e produtos nas redes sociais tem em média 39 anos[11], ou seja, faixa etária adulta e que possui poder aquisitivo. São as mesmas pessoas que vão não apenas opinar sobre o produto, mas também contaminar, através de compartilhamentos, *likes*, atualizações, comentários, mensagens privadas, toda a rede de contatos. Segundo Silva (2016, pág.8), porém, existem ainda mais desafios ao se tratar de *tweets*:

> "Além dos desafios que os sistemas de análise de sentimento que lidam com textos tradicionais enfrentam, a análise de sentimento no Twitter tem que lidar com dificuldades adicionais: tamanho do texto, variação na ortografia, esparsidade dos dados, definição de contexto e negação, são alguns dos muitos desafios [...]"

Um algoritmo pode ser definido como uma sequência de regras que serão aplicadas a um problema para que seja encontrada uma solução. Segundo o dicionário Michaelis, encontramos as definições[12] (grifo nosso):

- Processo de cálculo que, por meio de uma sequência finita de regras, raciocínios e operações, aplicada a um número finito de dados, leva à resolução de grupos análogos de problemas.
- Operação ou processo de cálculo; sequência de etapas articuladas que produz a solução de um problema; procedimento sequenciado que leva ao cumprimento de uma tarefa.
- Conjunto das regras de operação (conjunto de raciocínios) cuja aplicação **permite resolver um problema enunciado por meio de um número finito de operações**; pode ser traduzido em um programa executado por um computador, detectável nos mecanismos gramaticais de uma língua ou no sistema de

[11] Silva, 2016. Página 5
[12] https://michaelis.uol.com.br/moderno-portugues/busca/portugues-brasileiro/algoritmo/

procedimentos racionais finito, utilizado em outras ciências, para resolução de problemas semelhantes.

- Conjunto de regras e operações e procedimentos, definidos e ordenados usados na solução de um problema, ou de classe de problemas, em um número finito de etapas.

Para esse trabalho vamos utilizar dois algoritmos para comparação: NLTK e AFINN[13], disponível em inglês e na linguagem Python. Inicialmente testamos os da *Meaningcloud* e *Aylien* no software RapidMiner, porém, por motivos técnicos, tivemos que alterar a ferramenta de análise. Para simplificar, listamos as características de score de cada um:

Tabela 5 – Comparativo entre algoritmos

Algoritmo	Score	Plataforma
Meaningcloud	P+, P, Neu, N, N+, None	RapidMiner
Aylien	Valores entre 0 e 1	RapidMiner
AFINN	Valores entre + 5 e -1	Python
NLTK	Média entre positivo, negativo e neutro	Python

Segundo a empresa *Meaningcloud*, a análise funciona identificando no documento as relações entre entidades e frases para que seja possível determinar o sentimento do texto por completo. (tradução e aspas nossa):

> "Análise de Sentimentos ao nível de atributos. Nosso API de análise de sentimentos realiza uma análise de sentimentos detalhada e "multilinguagens" em informações de fontes distintas. O texto é analisado para determinar se ele expressa um sentimento positivo, neutro ou negativo (ou se é impossível detectar). Para tal, as frases individuais são identificadas e a relação entre elas avaliada, o que resulta em um valor global da polaridade do texto como um todo. Além da polaridade global e local, a API utiliza técnicas de processamento de linguagem natural avançadas para detectar a polaridade associada com ambas as entidades e os conceitos do texto. Ele também permite que usuários detectem a polaridade de entidades e conceitos que eles mesmos definem, o que faz com esta ferramenta seja aplicável em qualquer cenário."14

Mesmo após o tratamento dedicado na etapa de pré-processamento, atuar em 100% dos caracteres não é viável, visto que língua, seja ela escrita ou falada, é

[13] Apenas nos idiomas inglês e dinamarquês. Disponível em: <https://pypi.org/project/afinn/>.
[14] Disponível em: <https://www.meaningcloud.com/products/sentiment-analysis>

viva e cheia de coloquialismos. No tratamento de *tweets*, também como já vimos, podemos encontrar endereços de sites, fotos, GIF's e vídeos, gírias, apelidos, entre outros. Se casos assim não conseguirmos filtrar na etapa de pré-processamento, a árvore tentará classificar de acordo com as regras do dicionário da língua escolhida (português).

4.2 O projeto em Python

Ao iniciar esse estudo, pensávamos em utilizar para coleta, pré-processamento e mineração o software RapidMiner. Porém, ao longo dos estudos, identificamos que a exigência computacional do mesmo estava impossibilitando obtermos os resultados necessários para posterior análise. Logo, além de buscar outra alternativa que realizasse a análise textual, tivemos que nos adaptar ao novo cenário. Buscamos algoritmos que fizessem a análise de sentimentos em Python, pois os que estavam disponíveis anteriormente não trabalham com essa linguagem de programação. Também tivemos que alterar o idioma, pois não encontramos bibliotecas e algoritmos confiáveis e pré treinados a língua portuguesa.

Porém, como a análise de textos como um todo possui bastante documentação[15], conseguimos dar continuidade ao objetivo sem que alterássemos o objeto: como está sendo avaliado o personagem Harry Potter no Twitter. Faremos comentários, a seguir, dos macroprocessos que foram baseados no modelo proposto por Aranha no capítulo 3 e que acabamos por utilizar nesse estudo:

- A **coleta de dados** foi feita a partir de um segundo *script* que buscou, a base diretamente no Twitter
- No **pré-processamento** utilizamos as técnicas de *tokenização*, *stemming* e *stopwords*, além de retirar caracteres especiais, excesso de espaços, links de sites.
- Na **mineração** fizemos, a todo momento, a comparação entre 2 algoritmos, AFINN e NLTK, pois buscamos medir a qualidade do pré-

[15] Dispoível em: <https://rapidminer.com/>

processamento e a acurácia do modelo. Quanto mais próximos os dois algoritmos, para o nosso entendimento, melhor acurado ficou o modelo

- Na **análise**, verificamos os gráficos gerados pela biblioteca do *matplotlib*[16] e avaliamos, baseado nos resultados, se o personagem bem avaliado ou não.

[16] Disponível em: <https://matplotlib.org/>.

5. CONCLUSÃO

Como foi apresentado, discutido e analisado nos capítulos anteriores, o campo da Análise de Sentimentos textual é uma ferramenta de extrema importância para qualquer um que busca entender como são avaliados produtos, pessoas, personagens, acontecimentos, fatos, etc. Buscamos apresentar aqui termos que auxiliam nessa análise, como a implementação de um modelo de aprendizado de máquinas supervisionado. Mesmo com a adversidade (mudança de software de análise) que tivemos ao longo desse período, conseguimos experimentar como podemos montar, inferir, higienizar, ler, analisar, dados no qual queremos extrair respostas.

Porém, precisamos constantemente nos atentar ao espaço e tempo que está sendo produzido nosso banco de análise, pois eventos internos e externos a ele podem alterar o resultado, até mesmo nos apresentando um falso positivo. Optamos por pesquisar o personagem fictício Harry Potter porque sabíamos que a autora, J K Rowling, ao longo do mês de junho e julho, estava passando por uma baixa popularidade nas redes, após fazer comentários vistos como preconceituosos, transfóbicos e racistas pela comunidade – tendo tido até registros de vandalismos em seus monumentos pelo Mundo[17] -. O termo "cancelada", utilizado atualmente para excluir personalidades das redes sociais, foi veiculado à autora diversas vezes, inclusive em matérias de sites especializados em cultura jovem[18].

Sabemos também que lidamos aqui com um repositório orgânico, vivo, e que está constantemente em mutação nas expressões linguísticas de diferentes grupos sociais, empregando termos e vocabulários próprios do ambiente, como os neologismos e *emojis*, comentados no capítulo 6. Sendo assim, identificamos ao longo das pesquisas que poderíamos não obter 100% de acurácia na classificação das emoções verificadas. Ao analisar textos com padrões linguísticos definidos e da mesma fonte, com tesauros e dicionários próprios, ou blogs e matérias de jornais de

---

[17] Disponível em: <https://brasil.elpais.com/cultura/2020-07-12/ascensao-e-queda-de-j-k-rowling.html> e <https://revistamonet.globo.com/Noticias/noticia/2020/07/monumento-com-marca-das-maos-de-jk-rowling-e-vandalizado-com-tinta-e-bandeira-do-orgulho-trans.html>

[18] Disponível em: <https://www.omelete.com.br/quadrinhos/j-k-rowling-transfobia-entenda-polemica>.

um mesmo jornalista, encontraríamos, talvez, um ambiente que precisaria de menos atuação no pré-processamento – parte mais custosa do modelo -.

Concluímos que, apesar de as adversidades com relação à autora, o personagem continua avaliado de forma mais positiva que negativa, levando em conta apenas os 4, 5 mil *tweets* avaliados. Nas imagens a seguir, conseguimos identificar similaridades entre os gráficos, o que nos levou a crer que, independente do algoritmo e qual filtro utilizamos, o resultado se firma mais positivo que negativo.

No primeiro caso, buscamos a quantidade de classes (positivo, negativo e neutro) por região, no caso os Estados Unidos. Apesar de haver diferenças quando verificamos apenas os neutros, os dois algoritmos conseguiram identificar uma maior quantidade de *tweets* positivos e mesmo com uma menor avaliação neutra no NLTK, não foi apresentado aumento no negativo. Esse cenário persistiu em diversos gráficos. Na figura 10, apesar de os dois apresentarem *range* de avalição diferente, nota-se que existe uma tendência nas curvas. Na figura 11, avaliamos o número de *tweets* por classe, onde também notamos o mesmo padrão da figura 9.

Na análise do gráfico 12, onde aparecem os *tweets* ao longo do tempo (entre 31 de maio e 20 de julho), sendo a linha azul AFINN e a laranja NLTK, também notamos as mesmas tendências de curvas, mesmo que a discrepância no tamanho da curva seja perceptiva. O que importa aqui é notar que o termo não se manteve constante, tendo nas primeiras semanas de junho e julho a maior queda na popularidade.

Figura 9 – Classe por região

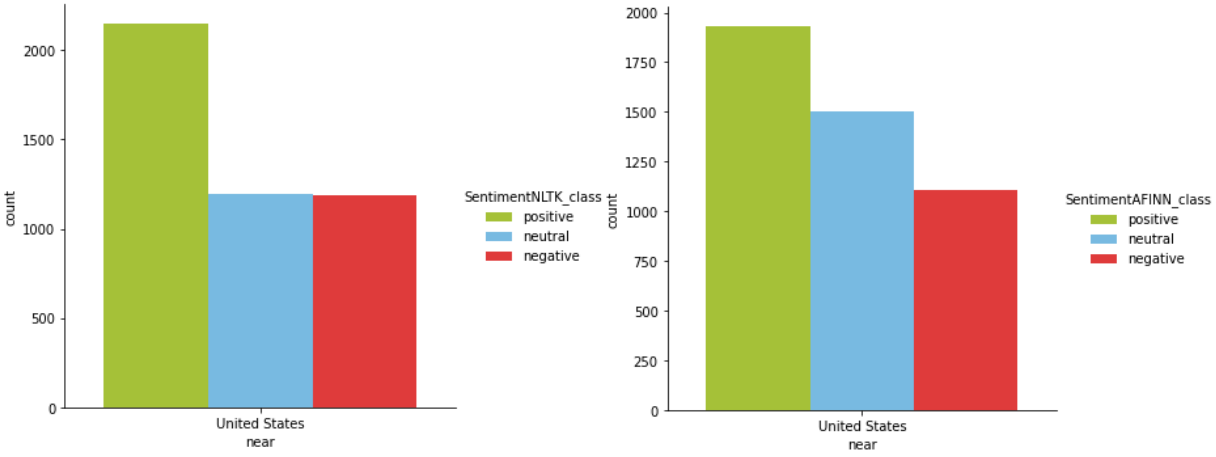


Figura 10 - Polaridade

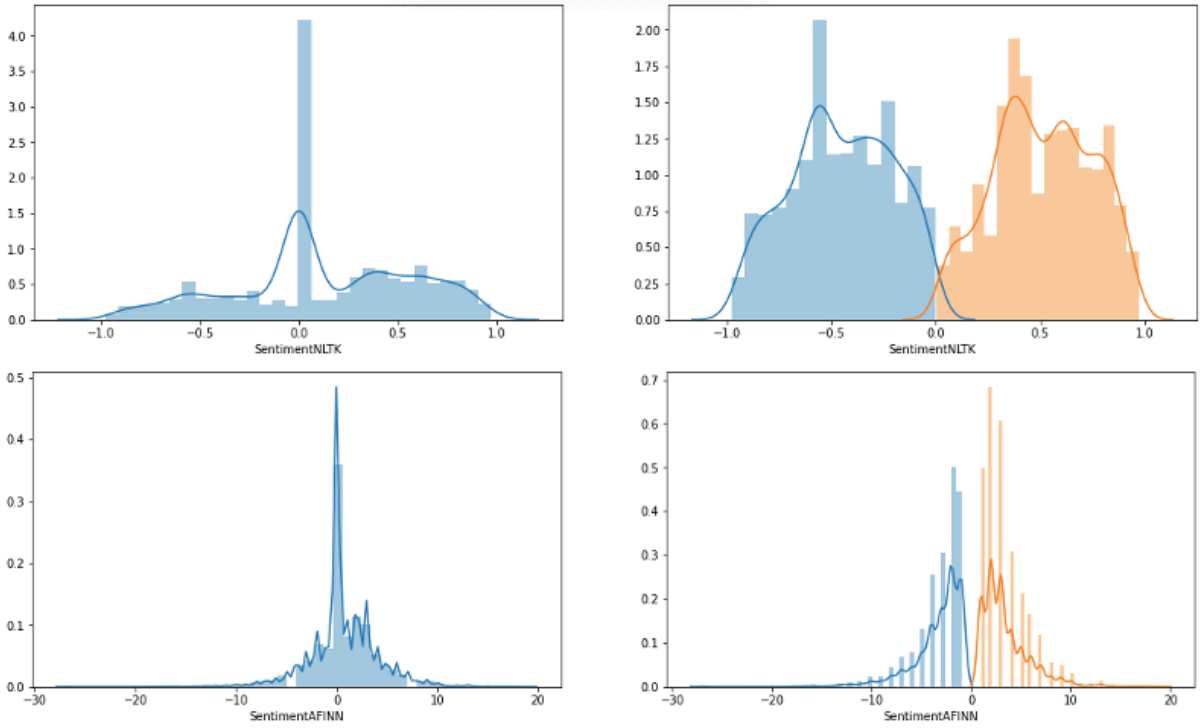


Figura 11- Número de *tweets* por classe

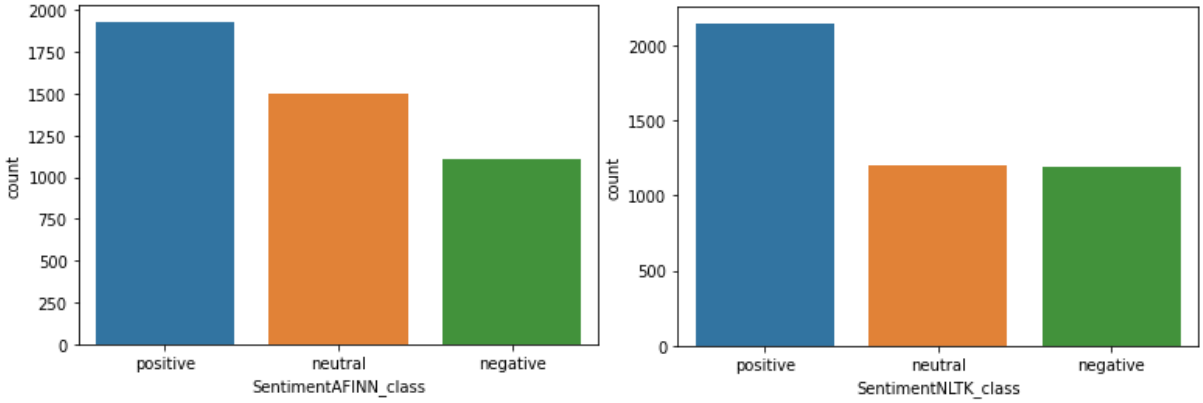
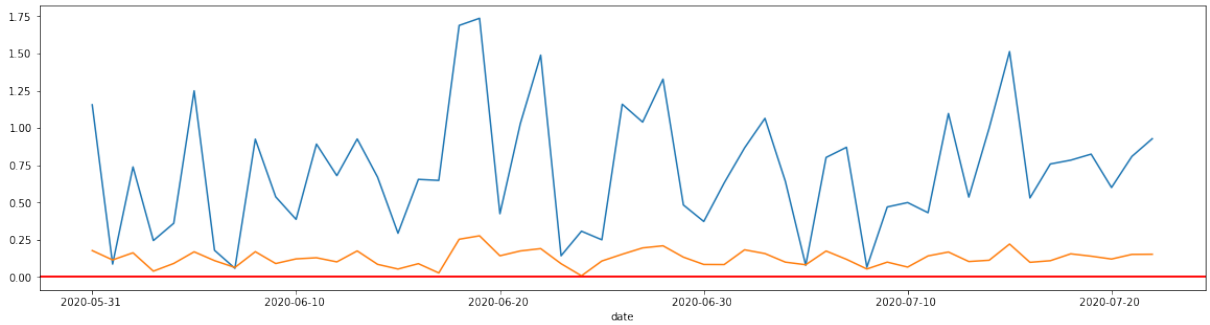


Figura 12 - Classe por data



REFERÊNCIAS BIBLIOGRÁFICAS

ALMEIDA Mariz, Anna Carla. RANGEL, Thayron Rodrigues. (organizadores). **Arquivologia: temas centrais em uma abordagem introdutória**. 1ª edição. Rio de Janeiro: FGV, 2020.

ARANHA, Christian Nunes. **Uma Abordagem de Pré-Processamento Automático para Mineração de Textos em português: Sob o Enfoque da Inteligência Computacional.** Pág. 60, 70-71. Tese (Doutorado em Engenharia Elétrica). PUC-Rio, Rio de Janeiro – RJ. 2007.

ALVARES, Reinaldo Viana. **Algoritmos de Stemming e o estudo de proteomas**. Pág. 9. Tese (Doutorado em Engenharia de sistemas e Computação). UFRJ, Rio de Janeiro – RJ. 2014.

BARREIRA, Rafael Gonçalves. **Análise de sentimentos com Rapidminer**. Pág. 16. Tese (Bacharel em Sistemas de Informação). Centro Universitário Luterano de Palmas. Palmas – TO. 2013.

BREIMAN, L. **Random forests**. 2001. Pág. 5-32. *apud* OSHIRO, Thais. **Uma abordagem para a construção de uma única árvore a partir de uma *Random Forest* para classificação de bases de expressão gênica**. Pág. 17. São Paulo, Universidade de São Paulo. 2013.

BRINGING HOGWARTS TO YOU. Disponível em: <https://www.wizardingworld.com/collections/harry-potter-at-home>. Acessado em: 23 e 27 de julho de 2020.

BRITO, Edeleon Marcelo Nunes. **Mineração de Textos: Detecção automática de sentimentos em comentários nas mídias sociais**. Pág. 17-18, 27. Tese (Mestrado em Sistemas de Informação e Gestão do Conhecimento) - Universidade Fundação Mineira de Educação e Cultura — FUMEC, Belo Horizonte – MG, 2017.

CANHISARES, Mariana. **J.K. Rowling e transfobia: entenda a polêmica**. Disponível em: <https://www.omelete.com.br/quadrinhos/j-k-rowling-transfobia-entenda-polemica>. Acessado em 27 de julho de 2020

COSTA E SILVA, L. da; TSUNODA, D. F.; DESLANDES, V. **Mineração de dados: busca de conhecimento sobre a evolução do canto da família *Thamnophilidae***. **AtoZ**, Curitiba, v. 1, n. 1, p. 61-70, jan./jun. 2011. Disponível em:<www.atoz.ufpr.br>. Acesso em 29 de junho de 2020.

DICIONÁRIO DICIO. **Termo: opinião**. Disponível em: <https://www.dicio.com.br/opiniao/>. Acesso em: 24 de junho de 2020.

DICIONÁRIO MICHAELIS. **Termo: algoritmo**. Disponível em: <https://michaelis.uol.com.br/moderno-portugues/busca/portugues-brasileiro/>. Acesso em: 8 de julho de 2020.

FILHO, Carlos Humberto Porto. **Técnicas de aprendizado não supervisionado baseadas no algoritmo da caminhada do turista**. Tese (Mestrado em Bioengenharia). Universidade de São Paulo, São Carlos – SP. 2017.

GAMBARO, Fernando. **Thermomether - Ed.9**. Disponível em: <https://br.kantar.com/mercado-e-pol%C3%ADtica/sa%C3%BAde-e-esporte/2020/thermometer-ed9/>. Acesso em: 24 de junho de 2020.

GONÇALVES, T. et al. **Analysing part-pf-speech for portuguese text classification. In:** ***Computacional Linguistics and Intelligent Text Processing.*** Springer, 2006. *apud* BRITO, Edeleon Marcelo Nunes. **Mineração de Textos: Detecção automática de sentimentos em comentários nas mídias sociais**. Belo Horizonte, Universidade FUMEC. 2017.

GUROVITZ, Hélio. **O que cerveja tem a ver com fraldas?**. Revista Exame. Disponível em: <https://exame.com/revista-exame/o-que-cerveja-tem-a-ver-com-fraldas-m0053931/>

JUNIOR, João Ribeiro Carrilho. **Desenvolvimento de uma Metodologia para Mineração de Textos**. 2007. Tese (Mestrado em Engenharia Elétrica). Universidade Pontifícia Universidade Católica do Rio de Janeiro - PUC-Rio, Rio de Janeiro – RJ**,** 2007.

KOHLER, Manoela. **Material apresentado na disciplina "Data Mining" para turma de BI Master**. PUC-Rio, Rio de Janeiro. 2018.

FERNANDES, Laura. **Ascensão e queda de J.K. Rowling**. Revista El Pais. Disponível em: https://brasil.elpais.com/cultura/2020-07-12/ascensao-e-queda-de-j-k-rowling.html. Acessado em 24 de julho de 2020.

LIU, B.; ZHANG, L. **A survey of opinion mining and sentiment analysis. Aggarwal, C. C.;** Zhai, C., editores *Mining Text Data*. Springer US *apud* SILVA, Nadia Felix Felipe da. **Análise de sentimentos em textos curtos provenientes de redes sociais**. USP - São Carlos, 2016.

LOPEZ, Alvaro Gustavo Talavera. **Controle Preditivo com Aprendizado por Reforço**
para Produção de Óleo em Poços Inteligentes. Tese (Mestrado em Engenharia Elétrica). PUC-Rio, Rio de Janeiro. 2010.

MAGALHÃES, Thiago. **Comentários no Twitter podem influenciar números de audiência.** Disponível em: <https://br.kantar.com/m%C3%ADdia/%C3%A1udio,-texto,-tv-e-v%C3%ADdeo/2016/abril-coment%C3%A1rios-no-twitter-podem-influenciar-n%C3%BAmeros-de-audi%C3%AAncia/>. Acesso em: 24 de junho de 2020.

MILLIGAN, G. W. **Clustering validation: Results and implications for applied analyses. Clustering and Classification**, Singapore: World Scientific, 1996. HENNIG, C.; LIAO, T. F. **Comparing latent class and dissimilarity based**

clustering for mixed type variables with application to social stratification (with discussion). Journal of the Royal Statistical Society, 2013. LUXBURG, U.; WILLIAMSON, R.; GUYON, I. **Clustering Science or art?** JMLR Workshop and Conference Proceedings, 2012. *apud* FILHO, Carlos Humberto Porto. **Técnicas de aprendizado não supervisionado baseadas no algoritmo da caminhada do turista.** Universidade de São Paulo. 2017.

MITCHELL, T. **Machine Learning**. McGraw Hill. 1997. New York, USA. *apud* BARREIRA, Rafael Gonçalves. **Análise de sentimentos com Rapidminer**. Palmas – TO. 2013.

NIELSEN, Finn ˚Arup. **A new ANEW: Evaluation of a word list for sentiment analysis in microblogs**. DTU Informatics, Technical University of Denmark. Lyngby, Denmark. 2011.

NETO, João Mendes de Oliveira. TONIN, Sávio Duarte. PRIETCH, Soraia Silva. **Processamento de Linguagem Natural e suas Aplicações Computacionais**. UFMT. Mato Grosso – MT. 2010.

OSHIRO, Thais Mayumi. **Uma abordagem para a construção de uma única árvore a partir de uma *Random Forest* para classificação de bases para expressão gênica**. Pág. 17. Tese (Mestrado em Bioinformática). Universidade de São Paulo, Ribeirão Preto – SP. 2013.

PEREIRA, Silvio do Lago. **Processamento de linguagem natural**. Disponível em: <https://www.ime.usp.br/~slago/IA-pln.pdf>. Acesso em: 24 de junho de 2020.

PRATI, Ronaldo Cristiano. **Novas abordagens em aprendizado de máquina para a geração de regras, classes desbalanceadas e ordenação de casos**. 2006. Tese (Doutorado em Ciências de Computação e Matemática Computacional) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2006. doi:10.11606/T.55.2006.tde-01092006-155445. Acesso em: 25 de junho de 2020.

REVISTA MONET. **Monumento com marcas das mãos de J.K Rowling é vandalizado com tinta e bandeira do orgulho trans**. Disponível em: <https://revistamonet.globo.com/Noticias/noticia/2020/07/monumento-com-marca-das-maos-de-jk-rowling-e-vandalizado-com-tinta-e-bandeira-do-orgulho-trans.html>. Acessados em 27 de julho de 2020.

ROSSI, Rafael Geraldeli. **Classificação automática de textos por meio de aprendizado de máquina baseado em redes**. 2015. Tese (Doutorado em Ciências de Computação e Matemática Computacional) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2015. doi:10.11606/T.55.2016.tde-05042016-105648. Acesso em: 2020-06-25.

SAYCE, David. **The Number of tweets per day in 2020**. Disponível em: <https://www.dsayce.com/social-media/tweets-day/>. Acesso em: 08 de julho de 2020.

SILVA, Nadia Felix Felipe. **Análise de sentimentos em textos curtos provenientes de redes sociais.** Pág. 4-5,8. Tese (Doutorado em Ciências de Computação e Matemática Computacional). USP - São Carlos, 2016.

SILVA, Renato Ramos da. **Aprendizado por reforço relacional para o controle de robôs sociáveis**. 2009. Dissertação (Mestrado em Ciências de Computação e Matemática Computacional) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2009. doi:10.11606/D.55.2009.tde-28052009-100159. Acesso em: 2020-06-24.

SITE CANALTECH. Estudo - **57% das empresas estão conectadas apenas para realizar vendas online**. Disponível em: <https://canaltech.com.br/redes-sociais/instagram-cresce-como-plataforma-de-vendas-nos-ultimos-meses-164417/>. Acesso em: 24 de junho de 2020.

SITE CUPONATION. **Saiba como as vendas online estão crescendo**. Disponível em: <https://www.cuponation.com.br/insights/vendasinstagram-2020>. Acesso em: 24 de junho de 2020.

SITE MATPLOTLIB. ***Documentation***. Disponível em: <https://matplotlib.org/>. Acessado em 27 de julho de 2020.

SITE MEANINGCLOUD. ***Sentiment Analysis***. Disponível em: <https://www.meaningcloud.com/products/sentiment-analysis>. Acesso em 8 de julho de 2020.

SITE PYPI. ***AFINN***. Disponível em: <https://pypi.org/project/afinn/>. Acesso em

SITE RAPIDMINER. ***Documentation***. Disponível em: <https://docs.rapidminer.com>. Acesso em: 25 de junho de 2020.

SNOWBALL. **Portuguese stemming algorithm**. Disponível em: <http://snowball.tartarus.org/algorithms/portuguese/stemmer.html>.Acesso em: 8 de julho de 2020.

SOARES, Matheus Victor Brum. **Aprendizado de máquina parcialmente supervisionado multidescrição para realimentação de relevância em recuperação de informação na WEB**. Pág. 15, 17. Tese (Mestrado em Ciências da Computação e Matemática Computacional). USP – São Carlos, SP, 2009.

APÊNDICE

- Máscara para geração de *wordcloud*. Nome: Harry_mask.png

- Script (Atualizado da base criada pelo Laboratório de Inteligência Artifical da PUC-Rio). Nome: HarryPotter

- Base retirada do Twitter. Nome: tweets.csv