INTERFACE TECNOLÓGICA

DO PROCESSAMENTO DE LINGUAGEM NATURAL À ANÁLISE DE SENTIMENTO

FROM NATURAL LANGUAGE PROCESSING TO SENTIMENT ANALYSIS

Edson Campos - edsonacampos@hotmail.com
Faculdade de Tecnologia de Taquaritinga (Fatec) – Taquaritinga – SP – Brasil

Mirela de Lima Piteli Picchi – mirela.piteli@fatectq.edu.br
Faculdade de Tecnologia de Taquaritinga (Fatec) – Taquaritinga – SP – Brasil

DOI: 10.31510/infa.v20i1.1667
Data de submissão: 20/03/2023
Data do aceite: 29/05/2023
Data da publicação: 30/06/2023

RESUMO

Neste artigo foi explanado sobre conceitos fundamentais de Processamento de Linguagem Natural e suas diversas ferramentas que tem se aprimoradas ao longo dos anos e se mostram a cada dia presente na rotina das pessoas. Realizou-se uma revisão e análise biográfica sobre Processamento de Linguagem Natural, Aprendizado de Máquina e Linguística Computacional, mostrando como a partir da análise textual pode-se chegar a complexos algoritmos de aprendizado de máquina, com o objetivo de conceituar e afunilar ao ponto de exemplificar em uma aplicação o conceito e o uso de Análise de Sentimento. A evolução do Processamento de Linguagem Natural torna necessária a evolução das ferramentas de aprendizado de máquina, pois são partes essenciais dentro de uma cadeia de eventos que vem se aprimorando com uma rapidez impressionante. Do conceito ao resultado final é possível verificar quão abrangente e repleto de ramificações é esse tema e, por meio de uma análise objetiva, constatar que o caminho para o futuro passa pelo aprendizado de máquina e a compreensão pela máquina da linguagem humana.

Palavras-chave: Processamento de Linguagem Natural. Aprendizado de Máquina. Análise de Sentimento.

ABSTRACT

This article explained fundamental concepts of Natural Language Processing and its various tools that have been improved over the years and are present every day in people's routines. A literature review and analysis was carried out on Natural Language Processing, Machine Learning and Computational Linguistics, showing how from textual analysis one can arrive at complex machine learning algorithms, aiming at conceptualizing and narrowing down to the point of exemplify the concept and application of Sentiment Analysis in an application. The evolution of natural language processing makes the evolution of machine learning tools necessary, as they are essential parts within a chain of events that has been improving with impressive speed. From the concept to the final result, it is possible to verify how comprehensive and full of ramifications this theme is and, through an objective analysis, verify

that the path to the future passes through machine learning and machine understanding of human language.

Keywords: Natural Language Processing. Machine Learning. Sentiment Analysis.

1 INTRODUÇÃO

A capacidade de se comunicar entre os seres humanos passou por muitas adaptações e aprimoramentos durante a história e hoje permite que as pessoas possam interagir entre si das mais diversas formas, algumas simples outras mais complexas. O aprimoramento da interpretação da linguagem humana se depara com mais uma fase da evolução, que é o processamento da linguagem humana por máquina, também conhecido por Processamento de Linguagem Natural.

O Processamento de Linguagem Natural é utilizado para compreender, interpretar e simular a linguagem natural das pessoas por meio de técnicas que incluem algoritmos, métodos estatísticos e aprendizado de máquina. O Processamento de Linguagem Natural não é algo novo, mas é uma tecnologia que vem evoluindo consideravelmente por conta do interesse de aperfeiçoamento da interação homem-máquina.

O objetivo deste trabalho é apresentar conceitos importantes sobre Processamento de Linguagem Natural e apresentar resultados de uma forma mais avançada de processamento e análise de dados que é a Análise de Sentimento. A Análise de Sentimento consiste em analisar um texto por uma máquina e classificar se o texto apresentado é positivo ou negativo. A Análise de Sentimento transforma-se diariamente em uma ferramenta avançada de inteligência industrial que colabora com a melhora de produtos e serviços e cria facilidades nas atividades das pessoas.

2 CONCEITUAÇÃO DO TEMA ABORDADO

Nesta seção serão apresentados conceitos essenciais de Processamento de Linguagem Natural e Análise de Sentimento abordando a abrangência e a aplicação dessas ferramentas.

2.1 Processamento de Linguagem Natural (PLN)

Pode-se dizer que o Processamento de Linguagem Natural é uma combinação entre Ciência da Computação, Inteligência Artificial e Linguística, com o objetivo de desenvolver ferramentas para gerar e compreender automaticamente a linguagem natural.

Para Titenok (2022), o Processamento de Linguagem Natural é um ramo da Inteligência Artificial que lida com a comunicação, sendo um método que permite que as máquinas possam criar e analisar a linguagem humana.

Para Vieira (2001), o que tornou possível a construção de sistemas com capacidade de reconhecer e produzir informações apresentadas em linguagem natural foi a exploração da relação entre linguística e informática, visando desenvolver modelos computacionais da cognição humana.

De acordo com Titenok (2022), O Processamento de Linguagem Natural foca nas estruturas semântica e gramatical presentes na linguagem. Essa característica possibilita lidar com as diversas variações da comunicação humana como imagens, discursos, texto, fala, entre outras. Por meio disso, é possível realizar muitas tarefas como traduções de qualidade entre diferentes línguas, análise dos sentimentos descritos na comunicação ou simplificar e tornar favorável a interação com os computadores para os seres humanos.

Ainda para Titenok (2022), os métodos de PLN são encontrados em lugares bem conhecidos como:

a) Mecanismos de busca: em pesquisas feitas no Google por exemplo, o retorno da pesquisa feita é aprimorado em cada pesquisa feita e adaptado para as pesquisas futuras, trazendo resultados mais objetivos em relação ao tema pesquisado;
b) Chatbots Inteligentes: o algoritmo PLN rodando em segundo plano por um gatilho especial para registrar que você precisa dele. Este gatilho aciona um programa de chatbot integrado ao seu canal de comunicação ou site e o orienta nos processos;
c) Verificação Ortográfica: aplicações de verificação ortográfica tem grandes bancos de dados de palavras, combinações de regras e palavras para que, quando uma palavra for inserida incorretamente, o sistema automaticamente sugira uma correção.;
d) Geração de Língua Natural, onde a partir de dados não linguísticos, seja uma planilha ou dados numéricos, pode-se utilizar uma aplicação de PLN para converter esses dados não-linguísticos em texto. Um grande exemplo para essa utilização é um robô desenvolvido por uma parceria entre a UFMG e a Poli/USP que coleta dados sobre o desmatamento da Amazônia, provindos do Instituto Nacional de Pesquisas Espaciais (INPE), convertendo dados numéricos, gerados via satélite, em textos que são publicados automaticamente no Twitter pelo perfil do robô o Da Mata

INTERFACE TECNOLÓGICA

Repórter. As figuras 1 e 2 apresentam, respectivamente, o perfil e um exemplo de postagem do referido robô.

Figura 1 - Da Mata Repórter

Fonte: Twitter (2023)

Figura 2 - Postagem Da Mata Repórter

Fonte: Twitter (2023)

2.2. Análise de Sentimento

O Processamento de Linguagem Natural envolve uma extensa gama de técnicas que se direcionam à aplicação de modelos e métodos analíticos computacionais ao conteúdo textual,

proporcionando ferramentas para análise de textos. Entre essas ferramentas, inclui Análise de Sentimentos, que pode ajudar os pesquisadores a explorar textos.

Para Saldaña (2018), a Análise de Sentimento busca quantificar e qualificar a intensidade emocional de palavras e frases dentro de um texto, levando em consideração, inclusive, o peso emocional de sinais linguísticos como pontuação ou mesmo emojis. As ferramentas de análise de sentimento que geralmente processam uma unidade de texto produzem pontuações e classificações quantitativas para indicar se o algoritmo considera que o texto transmite emoções positivas ou negativas.

De acordo com Almeida (2020), Análise de Sentimento se refere ao ato de analisar de forma automática algum fragmento textual, analisando o contexto de um determinado fragmento de texto e extraindo informações intrínsecas, e por vezes, subjetivas contidas nele.

2.2.1. Aprendizado de Máquina

Para entender melhor a automatização dessa ferramenta, é preciso entender o conceito de aprendizado de máquina.

Para Gonçalves (2021), Aprendizado de Máquina é um campo de estudo da área de Inteligência Artificial que ensina computadores a analisar e classificar padrões de dados, é uma programação de sistemas para assimilar dados e classificar informações complexas.

De acordo com Camelo (2017), projetar um computador capaz de aprender como um ser humano é uma tarefa há muito tempo explorada por pesquisadores que vem modelando o processo de aprendizagem em diversos algoritmos que hoje constituem a subárea da Inteligência Artificial chamada de Aprendizado de Máquina.

Um sistema de aprendizado de máquina tem a tarefa de analisar informações e identificar padrões para extrair um novo conhecimento.

2.2.2. Processamento de Textos

De acordo com Gonçalves (2021), a fase de processamento textual é uma das fases mais importantes de um projeto de aprendizado de máquina, pois ocorre a transformação dos dados. Essa transformação está relacionada com a conversão dos dados originais para formatos mais apropriados, a partir de normalização ou discretização de valores numéricos, a tokenização de palavras, entre outros, com o objetivo de diminuir a complexidade do processamento computacional.

Para Almeida (2021), o pré-processamento dos textos é crucial para convertê-los em estruturas interpretáveis pelos algoritmos de aprendizado de máquina, no entanto, para melhorar a performance dos mesmos é comum aplicar diversas técnicas de pré-processamento em conjunto. Algumas das mais utilizadas são:

a) Normalização de Texto: a normalização é um processo composto por várias etapas, sendo as mais comuns discutidas a seguir:

   I – Conversão para caixa baixa e remoção de pontuações: a conversão para caixa baixa é uma etapa recorrente em várias tarefas de PLN;

   II - Segmentação: também conhecido como tokenização e é o ato de segmentar um texto em palavras, este processo pode ou não manter pontuações e números;

   III – Remoção de stopwords: a remoção de palavras que aparecem com maior frequência que outras, mas oferece pouco valor semântico;

   IV – Lematização: que é a ação de definir a raiz de uma palavra por meio de análise morfológico e de vocabulário, e

   V – Stemming: onde o sufixo das palavras é removido, por algoritmos de lematização serem complexos, o stemming é uma ferramenta mais simples para o processo.

b) Representação de Texto: diversas aplicações que utilizam o Processamento de Linguagem Natural não conseguem aceitar entradas de texto puras, fazendo necessário que sejam vetorizadas essas entradas, as técnicas mais comuns de representação textual são:

   I – Bag-of-Word: também conhecido como mochila de palavras. É uma forma de representar um texto que descreve a ocorrência de palavras em um documento, levando em consideração um vocabulário de palavras conhecidas;

   II – Term-Frequency-Inverse Document Frequency(TF-IDF): este algoritmo reescala a frequência de palavras através do texto, penalizando palavras que tem maior frequência;

   III – Word-embedding: é uma técnica que aprende a representação dos significados das palavras, tornando possível fazer análise do contexto em um texto.

   c) Modelagem de linguagem: são modelos que atribuem probabilidades a sequencias de palavras, sendo uteis para identificar palavras relevantes no meio de entradas ambíguas.

3 PROCEDIMENTOS METODOLÓGICOS

Para exemplificar o que foi conceituado na seção anterior, foi criada uma aplicação para elucidar brevemente sobre PLN, precisamente Análise de Sentimento. Essa aplicação foi desenvolvida com a linguagem Python com suporte de algumas ferramentas.

3.1 NLTK

De acordo com a documentação NLTK (2023), NLTK é uma biblioteca para a construção de programas Python para trabalhar com dados de linguagem humana. Ele fornece interfaces fáceis de usar para mais de 50 corpora e recursos lexicais, como WordNet, juntamente com um conjunto de bibliotecas de processamento de texto para classificação, tokenização, lematização, marcação, análise e raciocínio semântico e um fórum de discussão ativo.

Para a aplicação foi utilizada a NLTK para tokenização e remoção das stopwords, importando-as da seguinte forma:

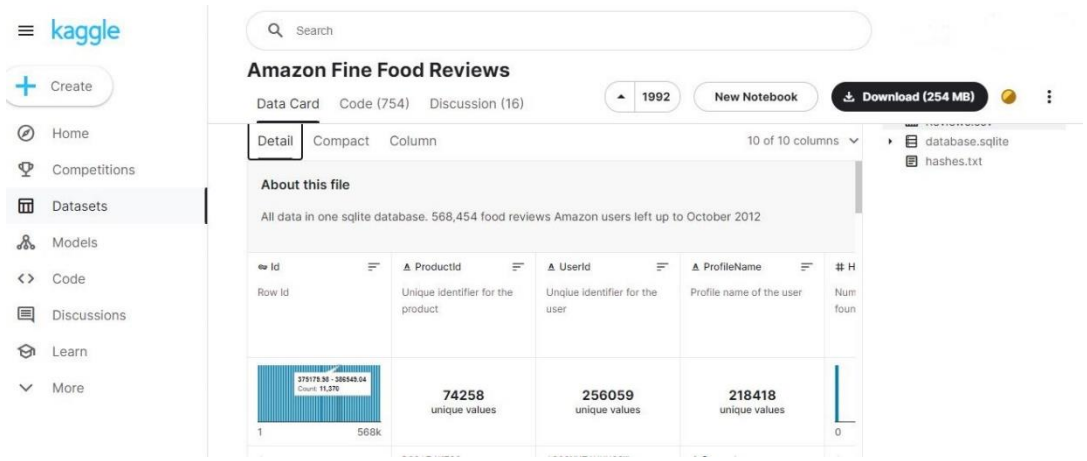
```
from nltk.tokenize import word_tokenize
from nltk.corpus import stopwords
```

3.2 Base de Dados

Para executar o processo de aprendizado de máquina é preciso ter um conjunto de dados de entrada para ser utilizado no treinamento da máquina para basear o aprendizado. Para esse processo, utilizamos uma base de dados pronta que está disponível na plataforma Kaggle (figura 3), que é uma grande comunidade de compartilhamento entre cientistas de dados. Além disso, utilizamos uma base de dados de reviews feitas por clientes da Amazon. O objetivo foi utilizar revisões positivas e negativas feitas por usuário e clientes que expressaram do que gostaram ou não. Dentro da base de dados analisamos duas possibilidades de revisões: positivas e negativas.

INTERFACE TECNOLÓGICA

Figura 3 - Plataforma Kaggle



Fonte: Kaggle (2023)

3.3. Naive Bayes

A vetorização dos textos é feita com base no algoritmo Naive Bayes (LEARN, 2023), é um conjunto de algoritmos de aprendizado supervisionados baseados na aplicação do teorema de Bayes com a suposição “ingênua” de independência condicional entre cada par de recursos dado o valor da variável de classe.

De acordo com Camelo (2017), o teorema de Bayes é uma fórmula que descreve como atualizar as probabilidades de hipóteses quando há evidências. Ele segue simplesmente dos axiomas da probabilidade condicional, mas possui uma ampla gama de uso.

3.4. N-Gram

De acordo com Andrade (2021), a técnica N-Gram é uma forma de conceder uma representação sintática aos modelos de extração de características que não mantém essa informação. Os N-Grams são conjuntos de sequência de palavras, onde N representa o número de tokens dentro de cada sequência.

Os conjuntos de extração podem ser feitos 1-Gram, 2 Gram, 3 Gram, e assim sucessivamente, sendo a extração com mais de um token com melhor resultado para obter mais informação sintática.

Exemplo de 1-Gram:

“Hoje eu irei à Fatec para estudar.”

No caso da sentença acima, o conjunto 1-Gram fica da seguinte forma:

{ “Hoje”, “eu”, “irei”, “à”, “Fatec”, “para”, “estudar” }.

Exemplo de 2-Gram:

“Hoje eu irei à Fatec para estudar.”

No caso da sentença acima, o conjunto 2-Gram fica da seguinte forma:

{ “Hoje eu”, “eu irei”, “irei à”, “à Fatec”, “Fatec para”, “para estudar” }.

4 RESULTADOS E DISCUSSÃO

Na aplicação desenvolvida, utilizou-se, nos vetorizadores, a técnica N-Gram, extraindo e analisando conjuntos com 1 e 4 tokens. Desse modo, o resultado exibe valores para as seguintes métricas: acurácia, precisão e medida F.

A métrica de acurácia trata o padrão da fração ou, de forma alternativa, da contagem de previsões corretas do método em questão. A métrica precisão trata a capacidade do método de não errar se uma entrada é positiva ou negativa. A métrica de medida F pode ser interpretada como uma média harmônica ponderada da precisão.

Para todas as métricas o resultado do treinamento de máquina indicará como sua melhor análise com um valor 1 e a pior análise com um valor 0.

A tabela 1 apresenta os resultados obtidos no treinamento da máquina:

Tabela 1 – Resultado Aprendizado de Máquina

	Acurácia	Precisão	MedidaF
1-Gram	0,865500	0,866362	0,865472
4-Gram	0,880000	0,881436	0,879947

Fonte: Elaborado pelos autores (2023)

A etapa seguinte foi testar o método de análise inserindo textos manualmente.

O primeiro teste é feito com sentenças simples: ‘Este produto é bom’. Em ambos os vetorizadores foi obtida a mesma análise: ‘Este texto é considerado positivo pelo classificador NaiveBayes’. Na sentença ‘Este produto é ruim’, foi obtida a análise: ‘Este texto é considerado negativo pelo classificador NaiveBayes’

No entanto, ao inserir um texto mais complexo, como: ‘Este produto não é perfeito’, obtém-se análises diferentes em cada classificador. O classificador com 1-Gram classifica o texto como ‘positivo’, enquanto o classificador com 4-Gram classifica o texto como ‘negativo’, isso ocorre pela análise assertiva criada na mochila de palavras no classificador com 4-Gram, dessa forma a análise feita a partir da junção dos termos denota a negatividade da sentença.

Com base nos testes feitos, concluímos que o classificador que apresenta melhor resultado é o que analisa o maior conjunto de tokens, tendo uma dimensão mais precisa dos textos analisados. Embora não seja possível garantir que a análise seja correta em todo e qualquer texto analisado, com quanto maior a base de dados usada para o treinamento melhor será a máquina treinada.

5 CONSIDERAÇÕES FINAIS

Dos conceitos básicos de Processamento de Linguagem Natural a aplicações mais avançadas no campo de Inteligência Artificial há uma gama enorme de ferramentas sendo desenvolvidas e usadas diariamente pelas pessoas no mundo todo. Tanto na área de Linguística como na área de Informática há estudos profundos para viabilizar e aperfeiçoar o entendimento da linguagem humana por máquinas. É indispensável ressaltar que, tão importante quanto desenvolver ferramentas de aprendizado de máquina, é preciso desenvolver a forma de se ensinar a linguagem textual, pois são partes essenciais desse processo.

REFERÊNCIAS

ALMEIDA, M. B. **Detecção Automática de Discurso de Ódio em Redes Sociais.** São José dos Campos, SP: UNIFESP, 2020.

ANDRADE, V. D. A. **Detecção Automática de Discurso de Ódio em Textos do Twitter.** Nova Iguaçu. RJ: Universidade Federal Rural do Rio de Janeiro, 2021.

CAMELO, F. A. B. **Detecção Automática de Discursos de Ódio em Comentários de Jornais Online**. Niterói, RJ: Universidade Federal Fluminense, 2017.

GONÇALVES, G. F. L. **Análise de Emoções em Tweets Relacionados à pandemia da Covid-19 no Estado do Rio de Janeiro.** Niterói, RJ: Universidade Federal Fluminense, 2021.

KAGGLE. **Amazon Reviews.** Disponível em: <https://www.kaggle.com/datasets/snap/amazon-fine-food-reviews>. Acessado em 20/03/2023.

LEARN, S. 1.2.2. **Naive Bayes. 2023.** Disponível em: <https://scikit-learn.org/stable/modules/naive_bayes.html>. Acessado em 16/02/2023.

SALDAÑA, Z. W. **Analise de Sentimento para Exploração de Dados.** 2018. Disponível em: <https://programminghistorian.org/pt/licoes/analise-sentimento-exploracao-dados>. Acessado em 15/02/2023.

TITENOK, Y. **Natural Language Processing vs Text Mining. 2022.** Disponível em: <https://sloboda-studio.com/blog/natural-language-processing-vs-text-mining>. Acessado em 20/02/2023.

TWITTER. **Da Mata Repórter.** Disponível em: <https://twitter.com/DaMataReporter>. Acessado em 19/02/2023.

VIEIRA, R.; LIMA, V.L.S. **Linguística computacional: princípios e aplicações**. In: As Tecnologias da Informação e a questão social: anais. Carlos Eduardo Ferreira (Ed.), SBC, Fortaleza, Ceará, Vol. 2, pp. 47-88, 2001.