UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"

Luiz Hemrique Cruz dos Santos
RA: 191251003

Projeto Análise de Dados – Introdução À Ciência de Dados

Graduação em Ciência da Computação
FCT - UNESP

Presidente Prudente
2024

Descrição

A base de dados escolhida para utilização neste projeto é composta de centenas de notícias no idioma inglês, marcadas como ‘Real’ ou ‘Fake’. Uma parte deste conjunto de dados é composto de notícias escritas e publicadas por fontes confiáveis, referentes à fatos ocorridos no cenário político dos Estados Unidos da América, enquanto a outra parte é composta de notícias falsas, sejam escritas e publicadas por websites que costumam disseminar esse tipo de notícia, quanto de redes sociais - como o X/Twitter, por exemplo.

É interessante o estudo desta base de dados, pois ela oferece uma oportunidade para o treinamento e teste de modelos de Aprendizado de Máquina – que serão feitos numa etapa posterior do projeto.

Exploração Inicial

A base de dados – que contém por volta de 10 mil entradas - é composta de dois atributos: uma coluna chamada ‘Text’, que corresponde ao texto completo da notícia; e outra coluna chamada ‘label’, que indica se a notícia é ’Real’ ou ‘Fake’.

A fim de facilitar a análise do conjunto de dados, foi realizada uma etapa de “limpeza” dos dados (em especial das entradas na coluna ‘Text’), utilizando métodos do módulo NTLK (Natural Language Toolkit).

Todas as etapas deste processo são necessárias para a utilização do conjunto de dados, como remoção de números, pontuação, remoção de palavras curtas, correção ortográfica etc. As entradas da coluna ‘label’ apenas foram transformadas de ‘Real’ e ‘Fake’ para 0 e 1, respectivamente.

Em específico, as etapas de *tokenização*, *stemming* (ou *lematização*) e *remoção de stop words* merecem alguma explicação.

Tokenização

Cada uma das entradas no conjunto de dados, isto é, cada uma das notícias, é dividida em palavras individuais, separando-as por espaço em branco e pontuações. Isso é útil para várias tarefas de processamento de linguagem natural, como análise exploratória, análise de sentimentos, identificação de tópicos etc.

Stemming / Lematização

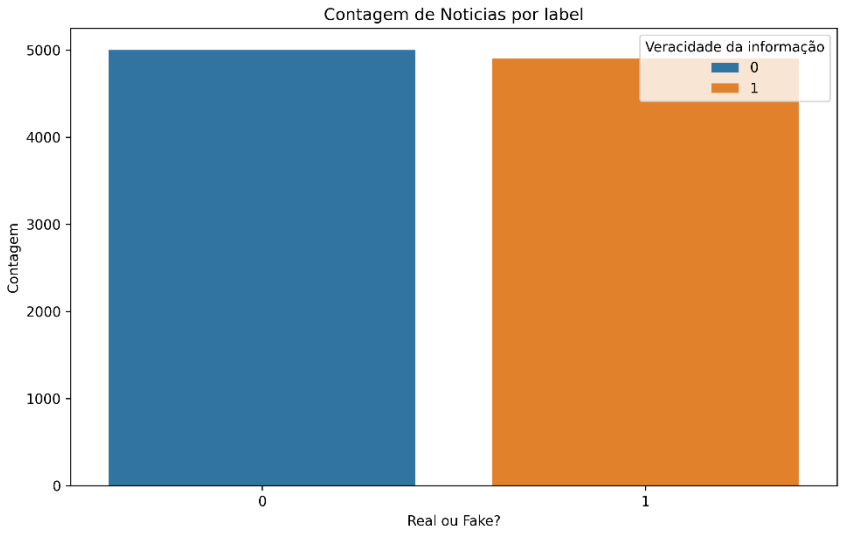
A etapa de *stemming* visa reduzir as palavras à sua forma raiz (stem) ou lema, que pode ajudar a reduzir o vocabulário e a normalizar o texto.

Remoção de stop words

Stop words são palavras consideradas como irrelevantes para a análise do conjunto de dados. São palavras usadas com frequência (como ‘the’, ‘a’, ‘an’, ‘in’ etc). A remoção destas palavras das entradas do conjunto de dados tem principalmente a vantagem de diminuir a quantidade de palavras que serão processadas durante a análise de dados.

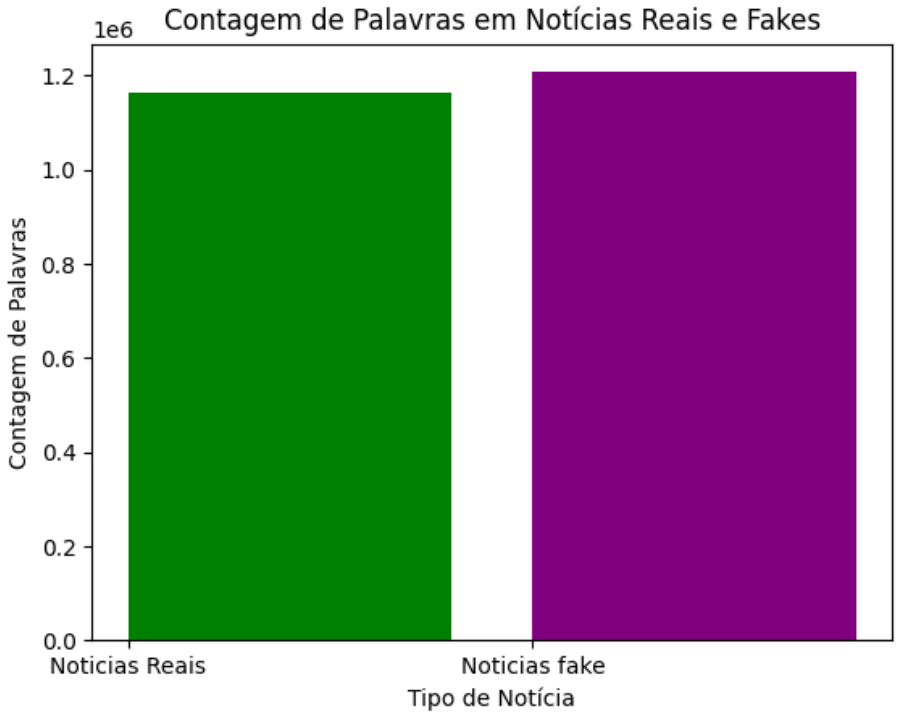
Contagem de Notícias por label (Real ou Fake)

A primeira análise feita foi em relação à quantidade de notícias para cada uma das categorias – Real ou Fake. Para o conjunto de notícias marcadas como Fake (ou 0) haviam 5000 valores, e para o conjunto de notícias marcadas como Real (ou 1) haviam 4900 valores. Estas contagens foram colocadas no seguinte histograma:



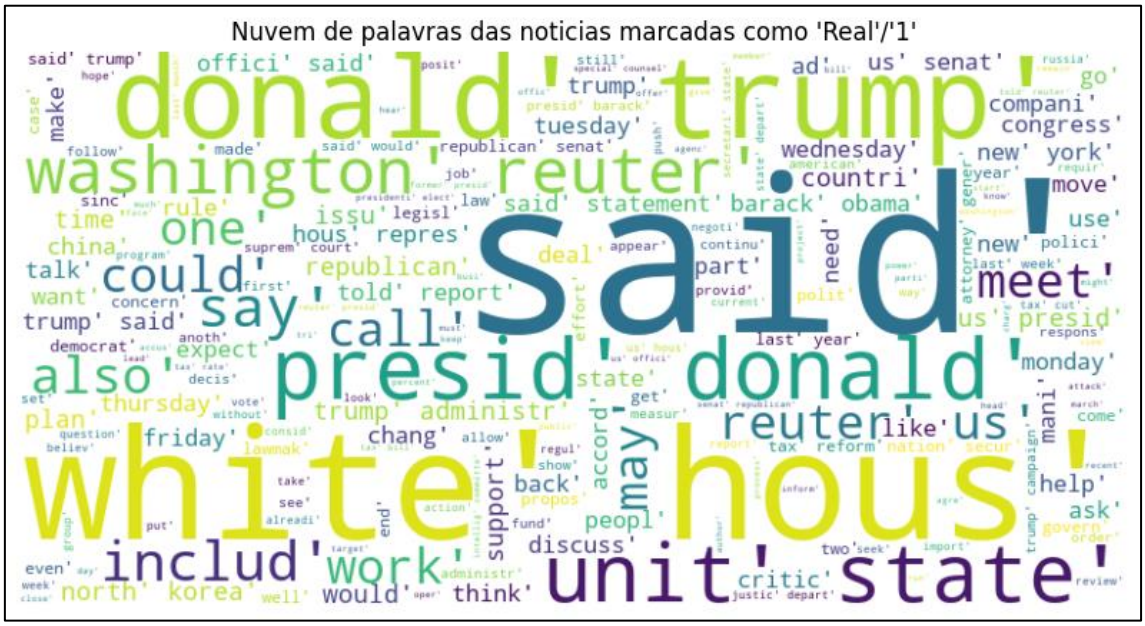
Contagem da quantidade de palavras por cada categoria de notícia

Outra análise inicial realizada foi a de contagem de palavras para cada uma das duas categorias de notícias. O número total de palavras em notícias marcadas como Real é 1162724; e o número total paras as notícias marcadas como Fake é de 1207771. Os valores estão representados no histograma a seguir.



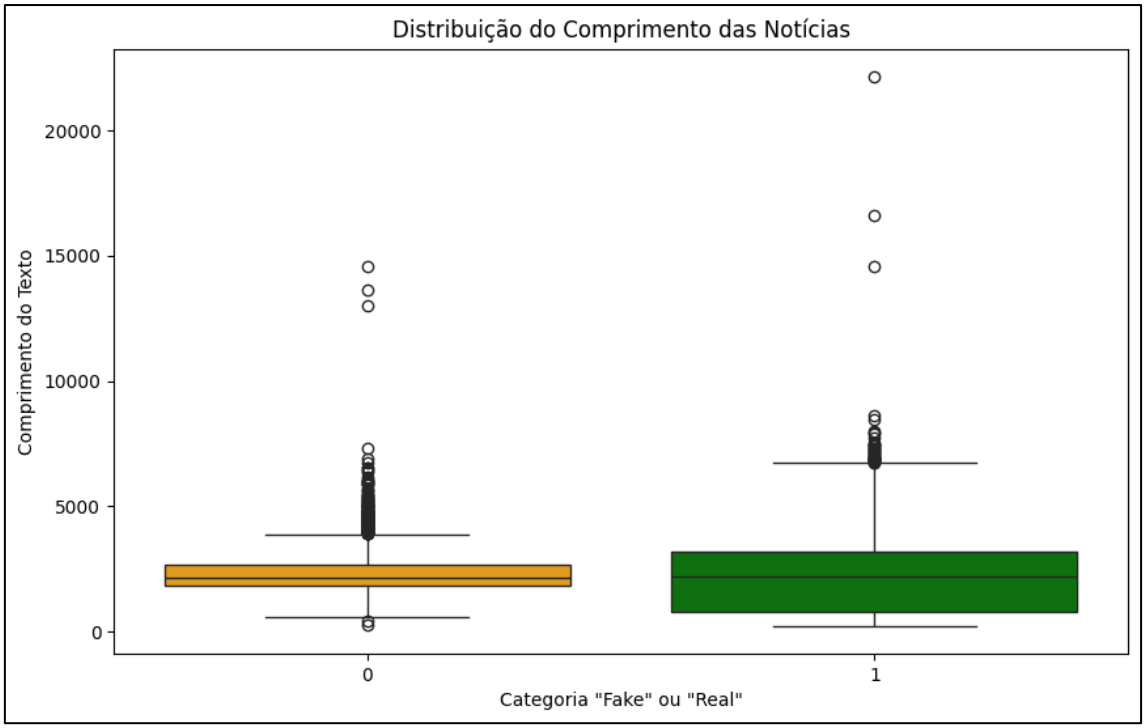
Nuvem de palavras

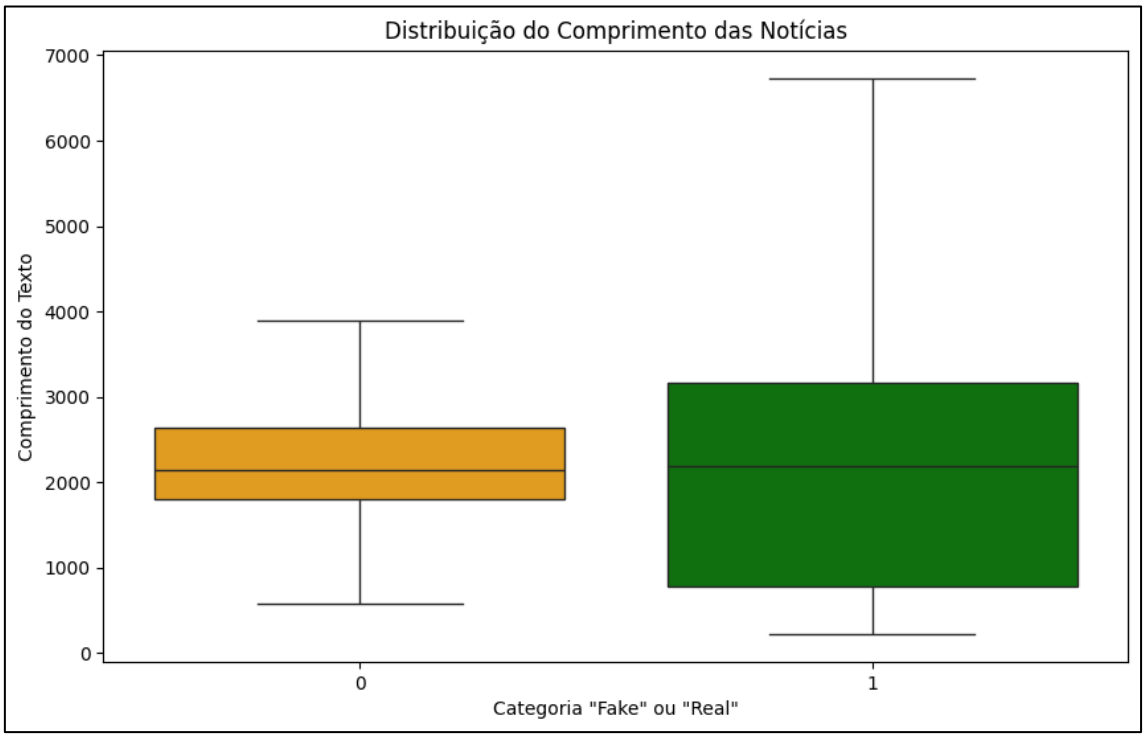
Uma maneira bastante útil de representar dados textuais é a nuvem de palavras, que representa a frequência ou importância de cada palavra dentro de um conjunto. A ideia é a de que palavras mais importantes ou frequentes apareçam maiores e com mais destaque, e as menos frequentes em menor tamanho. Dentro deste projeto, foram geradas duas nuvens de palavras: uma para o conjunto de notícias marcadas como Fake e outra pra o conjunto de notícias marcadas como Real.



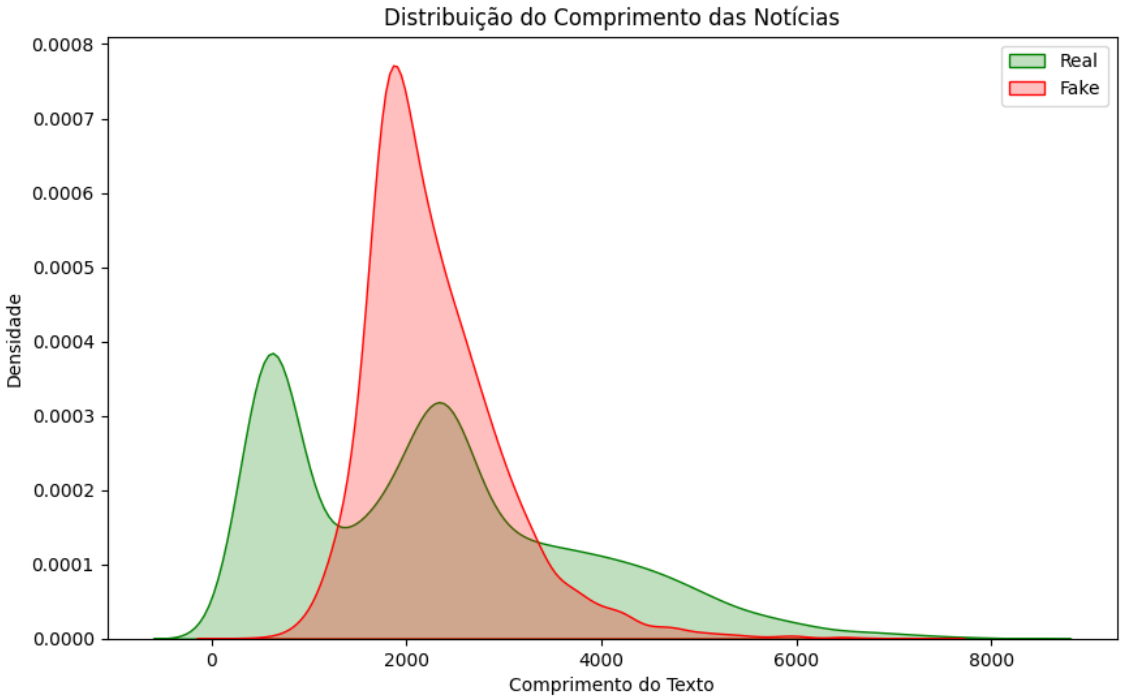
Comprimento das notícias

O comprimento das notícias, ou seja, a quantidade de palavras em cada notícia, foi representado em dois gráficos boxplot: um exibindo os outliers do conjunto, e outro sem os outliers.





Além disso, ainda relacionado ao comprimento das notícias, foi também criado um gráfico do tipo KDE (Kernel Density Estimate), que descreve a função de densidade de probabilidade das variáveis de dados contínuas ou não paramétricas.

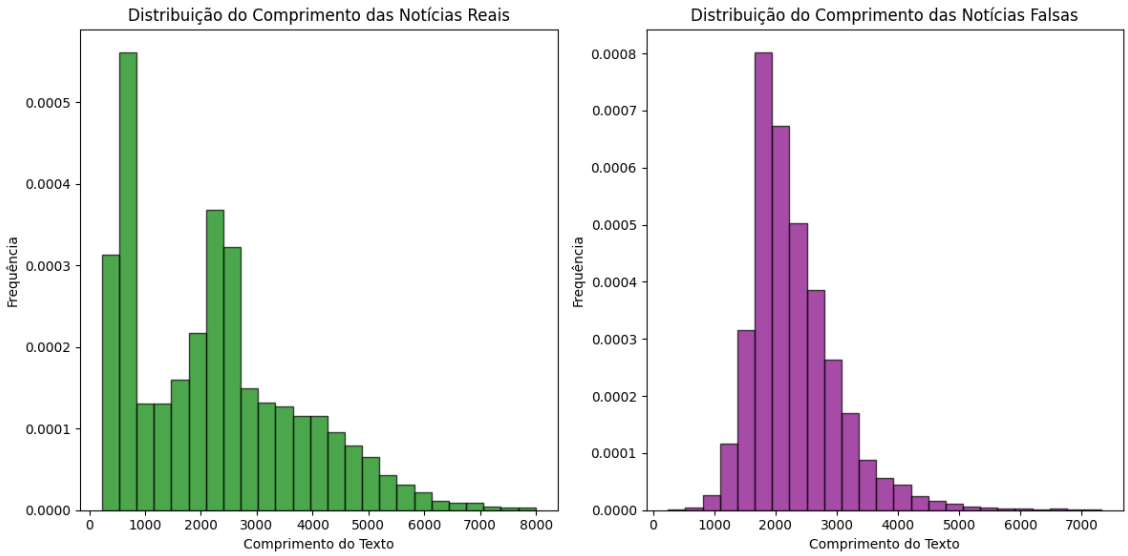


A distribuição para notícias marcadas como Real (em verde) tem um pico em um comprimento de texto menor em comparação com a distribuição para notícias marcadas como Fake (em vermelho), que tem seu pico em torno de 3000 palavras.

Ambas as distribuições diminuem à medida que o comprimento do texto aumenta, mas a distribuição para notícias falsas se estende mais do que a para notícias reais. Estas duas observações podem indicar que as notícias falsas tendem a ser mais longas do que as notícias reais neste conjunto de dados específica.

Comprimento do Texto

A quantidade de notícias por cada categoria (aqui chamada de “comprimento do texto”) também foi calculada, representada nos histogramas a seguir:



O gráfico mostra que as notícias marcadas como Real tendem a ter um comprimento menor, com uma frequência maior para textos mais curtos.

Outro detalhe facilmente visualizado no gráfico é que as notícias marcadas como Fake apresentam uma distribuição mais concentrada em torno de um comprimento específico de texto, indicando uma possível padronização no tamanho dessas notícias.

A comparação entre os dois gráficos sugere diferenças na maneira como as notícias marcadas como Real e Fake são escritas em termos de comprimento. Além disso, a frequência mais alta no gráfico de notícias marcadas como Fake sugere que há uma quantidade maior de notícias com um comprimento de texto específico, comparado às notícias marcadas como Real.

Frequência das palavras

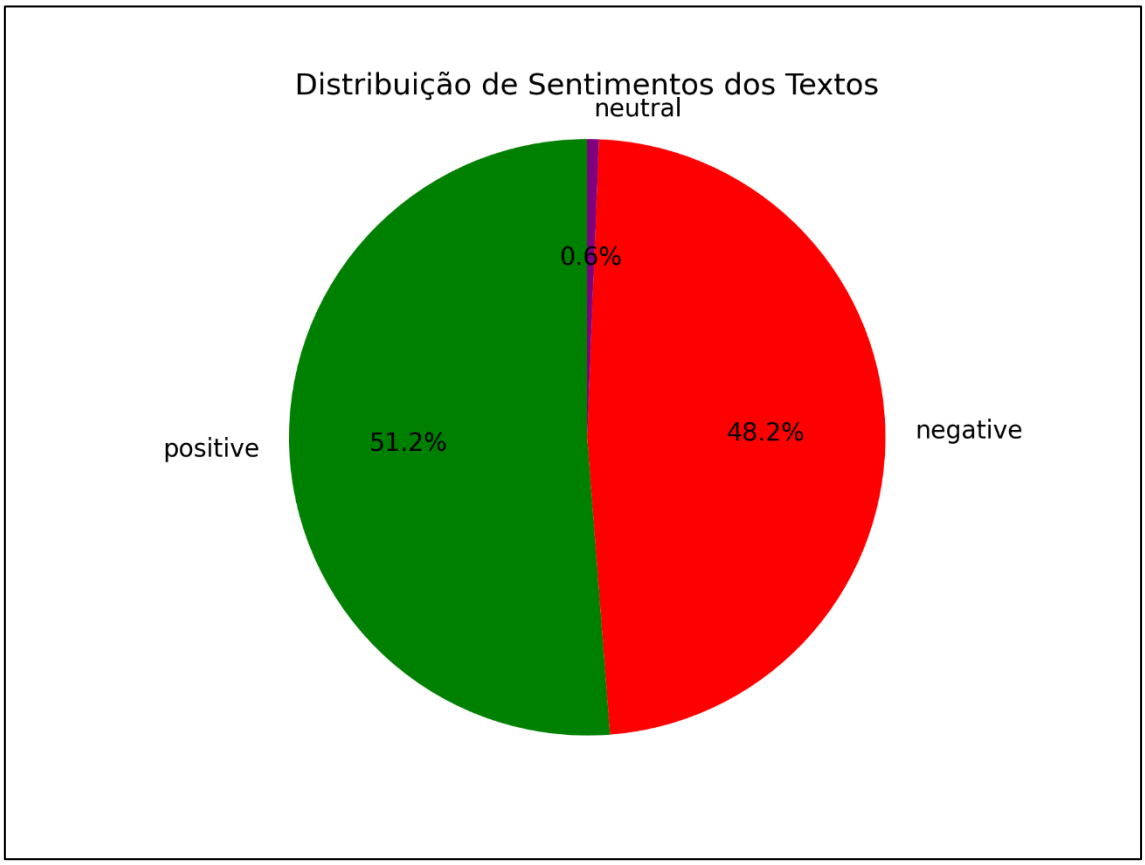
A limpeza de dados facilitou a contagem da frequência de palavras para cada categoria, permitindo extrair alguns dados básicos como média, variância, etc.

	CATEGORIA ‘REAL’	CATEGORIA ‘FAKE’
CONTAGEM	5767	5767
MÉDIA	184.4832668631871	190.08479278654414
DESVIO PADRÃO	666.146660	662.868445
MÍNIMO	6.0	6.0
25%	15.0	16.0
50%	40.0	45.0
75%	136.0	142.0
MÁXIMO	22084.0	36512.0
AMPLITUDE	22078.0	36506.0

Análise de Sentimento

A última análise realizada foi a ‘Análise de Sentimento’, que é uma técnica que busca determinar a orientação emocional ou a polaridade de palavras e frases de um texto. Ela pode identificar se o conteúdo expressa sentimentos positivos, negativos ou neutros.

A ferramenta utilizada nesta análise foi o VADER, e o resultado da análise de um texto é um conjunto de valores: *neg*, neu e pos. Esses três valores medem, respectivamente, a fração das pontuações ponderadas que se encaixam em cada categoria. Ele também calcula um valor *compound* (normalizado entre -1 e +1). Este valor descreve o efeito geral de todo o texto.



	VALORES
CONTAGEM	9900
MÉDIA	0.014551
DESVIO PADRÃO	0.80583
MÍNIMO	-0.999800
25%	-0.940225
50%	0.112600
75%	0.928800
MÁXIMO	0.999900
AMPLITUDE	0.0019

Ao analisar os resultados das estatísticas básicas, juntamente com a representação gráfica da distribuição de sentimento dos textos, fica claro que a maioria dos textos tem um sentimento positivo. Além disso, a média é próxima de zero, com um desvio padrão relativamente alto, o que sugere uma ampla variação nos sentimentos dos textos.