UNIVERSIDADE FEDERAL DO RIO GRANDE DO NORTE
BACHARELADO EM CIÊNCIAS E TECNOLOGIA

UTILIZAÇÃO DE NLP E SOM PARA ESTUDO DE ANALISE DE SENTIMENTOS NA REDE SOCIAL TWITTER

Componente Curricular: Tópicos Avançados de Informática I

Docente: Orivaldo Vieira de Santa Júnior

Discentes: 20200102204 Allyson Matheus Guedes de Oliveira

INTRODUÇÃO

A rede social Twitter nasceu com a proposta comum a todas as outras sociais, promover a interação dos usuários através de conversas, conteúdos midiáticos, notícias e tudo que for de interesse do público em comum. Atualmente é a principal rede social no que se diz respeito a mensagens e postagens rápidas, conteúdos atualizados com velocidade alta, e debates intensos sobre o que está acontecendo mundo a fora.

É muito comum ver pessoas utilizando o Twitter como um diário, fazendo uso da ferramenta de maneira a expor seus pensamentos e opiniões. Nisso, por se tratar de uma exposição que muitas vezes podem conter informações íntimas e quem sabe até de dados sensíveis.

Notou-se que seria interessante realizar um estudo e implementar uma ferramenta que faça a análise do que as pessoas, podendo elas serem cidadãos normais, ou até figuras públicas influentes, estão dizendo e pensando para prospectar suas futuras intenções e tomar medidas de acordo com o desejado, aferir quais foram os resultados e retirar as devidas conclusões.

DESENVOLVIMENTO

Toda a implementação foi realizada utilizando a linguagem de programação Python em diferentes ambientes de programação, via softwares instalados localmente como o Visual Studio Code, mas também através da ferramenta open source disponibilizada pelo Google chamada Colaboratory via serviço em nuvem.

O estudo foi realizado através da implementação de uma ferramenta baseada nos conceitos de Aprendizado de Máquina, ministrado na disciplina de Tópicos Avançados de Informática I. A primeira etapa consiste na utilização de ferramentas para a extração dos dados na rede social. A partir desta etapa utilizaremos os conhecimentos do processamento de linguagem natural (NLP) para o tratamento dos dados. Por fim, de modo analítico, escolhemos utilizar modelos de linguagem neural (SOM), de modo a aferir quais foram os resultados e retirar as devidas conclusões.

EXTRAÇÃO DOS DADOS

Os dados encontrados das contas são de acesso público quando a conta não tem o status privada. Porém seria humanamente impossível coletar cada publicação, uma por uma. Por isso existe a possibilidade da utilização de ferramentas que a própria empresa proporciona para extração ocorrer. Nos deparamos com duas possibilidades, utilizar um banco de dados já pronto e dispínel no site Kagle, que é uma plataforma especializada em hospedar diversos tipos de dados, ou realizar uma extração via aprendizado de máquina das contas específicas.

Para este trabalho, escolhemos as contas de twitter de duas das figuras públicas mais influentes do nosso país, o atual presidente de república e o ex-presidente, Lula e Bolsonaro, respectivamente.

O Twitter fornece uma plataforma para desenvolvedores, onde é disponibilizada uma API que oferece serviços e informações. Através dessa API, é possível obter os tweets públicos de qualquer perfil na rede. Para facilitar a interação com as informações da plataforma, o projeto utilizou a biblioteca Tweepy em Python, que oferece funções que simplificam a configuração e uso da API. O código fornecido como exemplo mostra como autenticar-se e baixar os tweets da linha do tempo da conta conectada, imprimindo cada um de seus textos no console.

```
# passa o Consumer Key e o Consumer Secret
auth = tweepy.OAuthHandler('', '')
# define o token de acesso
# passa o "Access Token" e o "Access Token Secret"
auth.set_access_token('', '')

# cria um objeto api
api = tweepy.API(auth)
```

Algoritmo: Exemplo de uso do algoritmo realizando a autenticação.

Por motivo de segurança, não vamos mostrar quais são as informações de autenticação explícitas, foi deixado apenas os espaços entre aspas (‘’) de modo ilustrativo. Com o uso do tweepy foi realizado o download das postagens dos usuários desejados e essas informações foram salvas em um arquivo no formato .csv.

	Unnamed: 0	Tweets
0	0	"teste"
1	1	RT @WhiteHouse: LIVE: President @realDonaldTru...
2	2	...love to have Mike Pompeo, Rick Perry, Mick ...
3	3	...lawyer has already stated that I did nothin...
4	4	The D.C. Wolves and Fake News Media are readin...
5	5	RT @DailyCaller: President @realDonaldTrump si...
6	6	RT @DailyCaller: Adam Schiff Challenger Jennif...
7	7	RT @DailyCaller: The Tide Is Turning Against D...
8	8	RT @DailyCaller: Poll: Independents Flip On Im...
9	9	When the Military rips down an old & badly...

Figura: print das primeiras linhas e colunas do arquivo salvo em formato .csv.

TRATAMENTOS DOS DADOS

Após obter os dados, é necessário realizar melhorias nos tweets e aplicá-los a um modelo de Processamento de Linguagem Natural (NLP) para extrair suas características. Optou-se por analisar apenas mensagens em inglês devido à disponibilidade de ferramentas adequadas.

A biblioteca NLTK (Natural Language Toolkit), escrita em Python, foi escolhida devido à sua popularidade e ampla gama de informações e exemplos disponíveis. Para a classificação do modelo NLP, escolheu-se o classificador Naive Bayes, que utiliza um modelo probabilístico de aprendizado baseado no teorema de Bayes. Para auxiliar o modelo, foi utilizada a ferramenta Vader Lexicon, que é uma ferramenta de análise de sentimentos baseada em regras e léxico, especialmente desenvolvida para capturar os sentimentos expressos nas mídias sociais.

Após aplicação do modelo obtivemos a seguinte tabelo com os dados já categorizados:

	Unnamed: 0	tweets	compound	negativos	neutro	positivo
0	0	"teste"	0.0000	0.000	1.000	0.000
1	1	RT @WhiteHouse: LIVE: President @realDonaldTru...	0.2960	0.000	0.804	0.196
2	2	...love to have Mike Pompeo, Rick Perry, Mick ...	-0.7177	0.140	0.860	0.000
3	3	...lawyer has already stated that I did nothin...	0.3724	0.000	0.943	0.057
4	4	The D.C. Wolves and Fake News Media are readin...	-0.7269	0.165	0.787	0.048
5	5	RT @DailyCaller: President @realDonaldTrump si...	0.0000	0.000	1.000	0.000
6	6	RT @DailyCaller: Adam Schiff Challenger Jennif...	-0.6249	0.242	0.693	0.065
7	7	RT @DailyCaller: The Tide Is Turning Against D...	0.0000	0.000	1.000	0.000
8	8	RT @DailyCaller: Poll: Independents Flip On Im...	0.0000	0.000	1.000	0.000
9	9	When the Military rips down an old & badly...	0.3707	0.100	0.759	0.141

Figura: Dados extraídos categorizados prontos para a análise.

ANÁLISE DOS DADOS

Com base nas informações obtidas na etapa anterior, em que os tweets foram categorizados de acordo com os sentimentos extraídos das palavras, optamos por utilizar uma rede neural auto-organizável chamada SOM (Self-Organizing Map). Essa escolha foi feita porque não temos as saídas desejadas e desejamos explorar as semelhanças entre os posts, a fim de agrupá-los em padrões semelhantes. A rede neural SOM permite uma análise e agrupamento eficiente dos dados com base em suas características e similaridades.

CONCLUSÃO

Após o estudo e implementação na NLP e o SOm com outro modelo de rede neural é possível obter bons resultados no reconhecimento de padrões em textos, conhecemos a riquíssima área de estudo do aprendizado de máquina e uma de suas melhores aplicações, ajudando a reconhecer padrões, problemáticos, ou que podem servir de análise, especificamente, no uso de rede sociais.

REFERÊNCIAS

ECT-INFO-ML. (2023). 2023_1.md. Recuperado de https://github.com/ect-info/ml/blob/master/2023_1.md

Kaggle. (s.d.). Lula-Bolsonaro Dataset. Recuperado de <https://www.kaggle.com/datasets/tleonel/lula-bolsonaro>

EACH-USP. (s.d.). Matrículas abertas para o PET-SI. Recuperado de <http://www.each.usp.br/petsi/jornal/?p=2747>

Redes Neurais Auto-Organizáveis - SOM. Recuperado de <http://aimotion.blogspot.com/2009/04/redes-neurais-auto-organizaveis-som.html>