SEUJAIR: Uma plataforma informatizada para análise autómatica da repercussão dos twittes do presidente do Brasil

Aluno: Fulano de Tal (CPF:***.***.***-**)
E-mail:flano@gamil.com ; Período da Graduação: VI
Orientador:

22 de maio de 2019

Resumo

Contexto: O uso das redes sociais vem crescendo nos últimos anos. Levando as pessoas a se exporem cada vez mais compartilhando conteúdo. Esse compartilhamento pode gerar uma serie de repercussões tanto negativas quanto positivas.

Problema: O uso por parte de figuras políticas pode gerar uma serie de repercussões. Onde o autor da postagem não chega ter compreensão se essas são positivas ou negativas gerando assim uma certa insegurança quanto as suas postagens.

Proposta: O presente pré-projeto visa o desenvolvimento de uma plataforma que seja capaz de realizar uma análise de forma automática da repercussão dos tweets do atual presidente do Brasil, Jair Bolsonaro, através mineração de textos.

Palavras-chaves:Mineração de Textos, Mineração de Dados, Tweets, Análise, Presidente, Jair Bolsonaro.

1 Introdução

A quantidade de dados informacionais disponíveis na internet tem aumentado alterando o modo como as pessoas buscam por notícias e informação. A internet criou a disponibilidade da troca e exposições de opiniões, fazendo com que as pessoas se mantenham cada vez mais conectadas em busca de conhecimento ou apenas para entretenimento.

Essa grande demanda de usuários navegando na internet levou posteriormente ao surgimento das redes sociais, que tem um forte componente de partilha e troca de informações sobre os mais diversos assuntos. No âmbito da informática as redes sociais são espaços virtuais onde grupos de pessoas ou empresas se relacionam através do envio de mensagens, da partilha de conteúdos, entre outros.

As mesmas tornaram-se ferramentas de comunicação muito populares entre os usuários da internet, como por exemplo o Twitter e o Facebook. Gerando uma grande quantidade de informações diárias, onde os usuários escrevem sobre os mais diversos conteúdos, desde a sua vida pessoal até assuntos relacionados a notícias.

O Twitter é uma das redes mais utilizadas no Brasil em 2019, e é uma importante fonte de opiniões sobre eventos e acontecimentos. O que levou a escolha dessa ferramenta para a pesquisa que será realizada. A rede teve um aumento significante na presença de usuários na passagem do ano de 2017 para o ano de 2018.

Essa ferramenta possibilita a comunicação entre as mais variadas pessoas. Nesse âmbito foi percebida a grande utilização do Twitter por parte do atual presidente do Brasil, Jair Bolsonaro. Onde foi observado a grande interação do mesmo na rede. Sabendo-se disso viu-se a necessidade de analisar se as postagens do atual presidente ocasionam um impacto positivo ou negativo perante a população que também utilizam a ferramenta. Desse modo identificando a repercussão das postagens do mesmo.

O restante deste pré-projeto está estruturado da seguinte forma. No capítulo 2 é abordado o referencial teórico através de conceitos essenciais ao entendimento do trabalho proposto. No capítulo 3 são abordados os trabalhos relacionados que apresentam uma comparação dos trabalhos mais relevantes relacionados a este pré-projeto. No capítulo 4 aborda-se a proposta, apresentada de forma mais detalhada como será feito para atingir os objetivos propostos, informando os métodos e ferramentas que serão utilizados. E por fim o capítulo 5 contendo o cronograma com os prazos para cada etapa do projeto seguido da bibliografia.

1.1 Objetivos Gerais e Específicos

O objetivo geral deste projeto é desenvolver uma plataforma informatizada que possa analisar de forma automática a repercussão das postagens do twitter do atual presidente do Brasil.

Para a concretização do objetivo geral, os objetivos específicos deste trabalho são:

1. Realizar um levantamento bibliográfico dos métodos a serem utilizados;
2. Implementar uma plataforma que possuirá ferramentas para mineração de textos e detecção de sentidos
3. Testar e interpretar os resultados obtidos a partir da plataforma em funcionamento

.

2 Referencial Teórico

Esta seção apresenta o referencial teórico que serve de sustentação para o entendimento do conteúdo abordado neste trabalho. Inicialmente é relatado conceito referente a mineração, de textos, e na sequência uma descrição das tecnologias necessárias para o desenvolvimento do presente trabalho.

2.1 Twitter

O Twitter surgiu em 2006 como uma rede social que permite aos seus utilizadores compartilhar infromação (tweets). Ao contrário de outras redes foca principalmente nas mensagens transmitidas. Onde tem como principal proposta levar o usuário a realizar uma publicação (tweet) sobre o que está fazendo ou o que está acontecendo ao seu redor.

È uma rede social de microblogging. Este tipo de blog nada mais é do que uma versão mini do blog original, porém com menos recursos e opções de interface. A média de caracteres é de 120 a 180 por post, os quais geralmente não ultrapassam três linhas. Como

a manutenção de blogs exige mais dedicação do blogueiro, muitos deles estão recorrendo ao microblogging para não deixar de lado o hobby de compartilhar suas opiniões ou discussões de temas do seu interesse.

É uma das redes sociais que têm mais utilizadores no mundo inteiro. E também, uma ferramenta de marketing imprescindível para todas as empresas, marcas, serviços ou produtos. O Twiter, serve para comunicar com a uma rede de seguidores, que vai construindo com o passar do tempo. Vai servir como um canal de comunicação entre a pessoa que cria a conta no Twitter e os seus seguidores.

2.2 Text Mining

Text mining ou mineração de textos é uma área que tem como principal objetivo extrair conhecimento implícito de grandes quantidades de textos escritos em linguagem natural. Sua definição é bastante semelhante àquela dada à mineração de dados, sendo a grande e importante diferença o meio utilizado para minerar: na data mining, a mineração é realizada em uma base de dados.

O processo de mineração de texto não é simples, mas possui pilares que direcionam qualquer projeto ao êxito: objetivo bem definido, profissionais qualificados, softwares e metodologia são a base do sucesso.

2.3 Tokenizer

Essa etapa consiste na separação de um texto em suas unidades mínimas, ou seja, deve-se separar cada palavra presente no texto, a pontuação, etc. Tais unidades mínimas são conhecidas como tokens. Existem abordagens que agrupam dois tokens para formar um token com significado agregado pois, por exemplo, o nome composto de uma entidade tem seu valor no seu conjunto de palavras, e não em cada palavra separadamente.

É importante perceber que o processo de tokenização pode parecer simples, mas sua dificuldade provém da separação de termos que não poderiam ser separados.

2.4 JSON

JSON (JavaScript Object Notation) é um formato leve de troca de dados. É fácil para humanos ler e escrever. É fácil para as máquinas analisar e gerar. É baseado em um subconjunto da Linguagem de Programação JavaScript , Padrão ECMA-262 3ª Edição - Dezembro de 1999 . JSON é um formato de texto completamente independente do idioma, mas utiliza convenções que são familiares aos programadores da família C de linguagens, incluindo C, C++, C#, Java, JavaScript, Perl, Python e muitos outros. Essas propriedades tornam o JSON uma linguagem de intercâmbio de dados ideal.

O JSON é construído em duas estruturas:

Uma coleção de pares nome/valor. Em vários idiomas, isso é realizado como um objeto , registro, estrutura, dicionário, tabela de hash, lista com chave ou matriz associativa. Uma lista ordenada de valores. Na maioria das linguagens, isso é realizado como uma matriz , vetor, lista ou sequência. Estas são estruturas de dados universais. Praticamente todas as modernas linguagens de programação as suportam de uma forma ou de outra. Faz sentido que um formato de dados que seja intercambiável com linguagens de programação também seja baseado nessas estrutu

2.5 OpenNPL

A biblioteca Apache OpenNLP é um kit de ferramentas baseado em aprendizado de máquina para o processamento de texto em linguagem natural.

O OpenNLP suporta as tarefas mais comuns da PNL, como tokenização , segmentação de frase , marcação de parte da fala , extração de entidades nomeadas , chunking , análise , detecção de idioma e resolução de referência .

2.6 Inteligência Artificial

Inteligência Artificial (IA) é um ramo da ciência da computação que se propõe a elaborar dispositivos que simulem a capacidade humana de raciocinar, perceber, tomar decisões e resolver problemas, enfim, a capacidade de ser inteligente.

Existente há décadas, esta área da ciência é grandemente impulsionada com o rápido desenvolvimento da informática e da computação, permitindo que novos elementos sejam rapidamente agregados à IA.

Iniciada dos anos 1940, a pesquisa em torno desta incipiente ciência eram desenvolvidas apenas para procurar encontrar novas funcionalidades para o computador, ainda em projeto. Com o advento da Segunda Guerra Mundial, surgiu também a necessidade de desenvolver a tecnologia para impulsionar a indústria bélica.

Com o passar do tempo, surgem várias linhas de estudo da IA, uma delas é a biológica, que estuda o desenvolvimento de conceitos que pretendiam imitar as redes neurais humanas. Na verdade, é nos anos 60 em que esta ciência recebe a alcunha de Inteligência Artificial e os pesquisadores da linha biológica acreditavam ser possível máquinas realizarem tarefas humanas complexas, como raciocinar.

2.7 JASON

JSON (JavaScript Object Notation) é um formato para intercâmbio de dados considerado simples tanto para a leitura e escrita humana quanto para a realização das mesmas coisas de maneira computacional. É definido em formato texto e está disponível para uso em várias linguagens, como C, C++, Java, Perl, Python, etc. No nosso caso utilizaremos java.

Sua constituição se dá em duas partes: uma sequência de pares nome/valor e uma sequência ordenada de valores. Essas são estruturas suportadas na maioria das linguagens de programação (objetos, registros, dicionários, etc, correspondem à sequência de pares nome/valor e vetores, listas, etc, correspondem à sequência ordenada de valores).

2.8 JAVA

Originalmente desenvolvida por uma equipe de desenvolvedores liderada por James Gosling na Sun Microsystems (atualmente de propriedade da Oracle) e lançada em 1995, o Java é uma linguagem de programação orientada a objetos que atualmente faz parte do núcleo da Plataforma Java.

A Orientação a Objetos, ou Programação Orientada a Objetos (POO), do inglês Object-Oriented Programming (OOP), é um tipo de paradigma de análise, para a programação de sistemas no qual todos os elementos inseridos são objetos. Foi uma das tentativas de trazer a programação para um nível de linguagem mais semelhante ao cotidiano.

O desenvolvedor é responsável por modelar o papel desempenhado pelos objetos e a interação entre eles. Por exemplo, em um sistema desenvolvido para uma padaria, existiriam objetos do tipo "Cliente"e objetos que simulam as ações que um cliente pode realizar.

No momento de seu desenvolvimento, os objetivos principais desejados para esta linguagem foram que ela deveria ser simples, orientada a objetos e de fácil aprendizagem não somente para programadores experientes.

3 Trabalhos Relacionados

É possível encontrar estudos conhecidos relacionados às técnicas e ferramentas que serão necessárias a este projeto na literatura disponível. A Tabela 1 lista alguns trabalhos com objetivos relacionado aos apresentados na proposta deste projeto. Para fator de comparação foram usadas as seguintes métricas: Objetivo, Objeto Analisado, Ferramentas e Técnicas Utilizadas. Cada uma estará comentada abaixo.

O trabalho desenvolvido por nascimento2012analise teve como objetivo concluir apartir dos sentimentos expostos nos twettes, se a população achou um determinado fato positivo ou negativo. Porém não apresentou uma coleta e associação de twittes de forma autmatizada.

Já o trabalho de cavalcanti12marcas realiza uma análise com relação ao nível de repercussão dos assuntos que as marcas Sky, Fiat e Bradesco, postam no Twitter. Más não exploram as relações dos usuário que postam (as empresas) e o público.

Na pesquisa de recuero2014discurso que teve como objetivo discutir o discurso dos protestos de junho de 2013 no Brasil a partir de um recorte específco de tweets. Não são apresentados meio de análise automatica de um perfil em particula, mas somente de marcações.

É possivel destacar também o trabalho de TEIXEIRA2011 que aplicou técnicas de Análise Sentimental para verificar se a informação existente em duas redes sociais (Facebook e Twitter) pode ser utilizada para estimar valores que podem vir a ser obtidos na comercialização de bens ou serviços a serem lançados no mercado. Pode-se observar que não se houve a preocupação com os perfis que serim análisados, más somente com a marcação de filme.

4 Proposta

Este pré-projeto tem como objetivo o desenvolvimento de uma ferramenta que deverá realizar a análise do twitter do atual presidente do Brasil. Para isso, a ferramenta realizará uma análise do perfil do mesmo através da mineração de texto.

Primeiramente será necessário realizar a extração, pré-processamento e armazenamento dos tweets do atual presidente Jair Bolsonaro e dos comentários, a extração será realizada com o uso de aplicações-cliente em linguagem java da Streaming API, visto permitir a obtenção de comentários mais antigos sem a necessidade de autenticação. O retorno dos resultados das duas redes sociais é feito em JSON (JavaScript Object Notation), um formato de fácil leitura e manipulação.

Após a obtenção de mensagens será realizado o seu pré-processamento, que altera a escrita informal, as abreviaturas e a ênfase dada às palavras através da repetição de

caracteres, bastante comuns nas redes sociais, mas que originam uma avaliação errada pelas técnicas habituais.

Em segundo caso será necessário realizar o processamento de linguagem natural que tem o objectivo de facilitar a posterior análise sentimental das mensagens. Para a implementação das técnicas de processamento de linguagem natural foi utilizada a ferramenta OpenNLP. Como explicado na secção anterior, existem diversas técnicas de processamento de linguagem natural. Neste projecto foram implementados dois processos considerados relevantes, tokenization e o POS Tagger (Part of Speech Tagger). O processo de tokenization realiza a divisão do texto em tokens, de modo a facilitar o tratamento posterior de cada token de forma independente.

Para a realização desta operação é necessário importar os modelos treinados da ferramenta OpenNLP, que contêm as regras necessárias para a correcta execução dos diversos métodos de processamento.

Em terceiro caso será realizada a análise sintática das mensagens, com a utilização do componente Phrase Parser da ferramenta OpenNLP. O Phrase Parser faz a análise das mensagens, identificando os vários assuntos presentes na frase e quais os elementos (sujeitos, adjectivos, entre outros) que se relacionam com esse mesmo assunto, fazendo o seu agrupamento.

E por último a análise sentimental que possibilita das mensagens extraídas que tem como resultado a obtenção da sua polaridade. Para a sua realização, todos os processos descritos anteriormente nesta secção de desenvolvimento são necessários, visto procederem ao tratamento das mensagens que serão agora analisadas de modo a ser obtida a sua polaridade.

Neste método é analisada uma mensagem de cada vez, sendo dada maior relevância aos substantivos, que expõem os assuntos da frase e aos adjectivos, que os caracterizam como positivos ou negativos.

A plataforma que será desenvolvida em java deverá possuir todas essas ferramentas para que seja possível realizar todos esses procedimentos.

4.1 Avaliação/Estudos de Caso

A avaliação dos resultados será feita a partir da acurácia, entre os dados obtidos na plataforma e a análise dos dados de forma manual. Inicialmente será feita uma execução de analise na plataforma, e depois será realizada uma análise manual para que seja possível realizar a avaliação.