UNIVERSIDADE FEDERAL DA FRONTEIRA SUL
CAMPUS DE CHAPECÓ
CURSO DE CIÊNCIA DA COMPUTAÇÃO

JOD FEDLET PIERRE

UTILIZAÇÃO DE MODELAGEM DE TÓPICOS PARA IDENTIFICAR OS ASSUNTOS MAIS DISCUTIDOS SOBRE O HAITI NAS REDES SOCIAIS

CHAPECÓ
2022

JOD FEDLET PIERRE

UTILIZAÇÃO DE MODELAGEM DE TÓPICOS PARA IDENTIFICAR OS ASSUNTOS MAIS DISCUTIDOS SOBRE O HAITI NAS REDES SOCIAIS

Trabalho de conclusão de curso apresentado como requisito para obtenção do grau de Bacharel em Ciência da Computação da Universidade Federal da Fronteira Sul.
Orientador: Prof. Dr. Denio Duarte

CHAPECÓ
2022

Pierre, Jod Fedlet

Utilização de modelagem de tópicos para identificar os assuntos mais discutidos sobre o Haiti nas redes sociais / Jod Fedlet Pierre. – 2022.

42 f.: il.

Orientador: Prof. Dr. Denio Duarte.

Trabalho de conclusão de curso (graduação) – Universidade Federal da Fronteira Sul, curso de Ciência da Computação, Chapecó, SC, 2022.

1. Modelagem de tópicos. 2. Twitter. 3. Haiti. 4. Bertopic. I. Duarte, Prof. Dr. Denio, orientador. II. Universidade Federal da Fronteira Sul. III. Título.

E-mail: fedletpierre15@gmail.com

JOD FEDLET PIERRE

UTILIZAÇÃO DE MODELAGEM DE TÓPICOS PARA IDENTIFICAR OS ASSUNTOS MAIS DISCUTIDOS SOBRE O HAITI NAS REDES SOCIAIS

Trabalho de conclusão de curso apresentado como requisito para obtenção do grau de Bacharel em Ciência da Computação da Universidade Federal da Fronteira Sul.

Orientador: Prof. Dr. Denio Duarte

Este trabalho de conclusão de curso foi defendido e aprovado pela banca avaliadora em: 29/3/2022.

BANCA AVALIADORA

Prof. Dr. Denio Duarte – UFFS

Prof. Dr. Felipe Grando – UFFS

Prof. Me. Geomar Schreiner – UFFS

RESUMO

Com o crescimento de acesso a internet da população mundial, grandes volumes de dados, principalmente no formato de textos, vêm sendo compartilhados diariamente. Seja por meio de redes sociais, blogs ou fóruns. Esses dados podem ser analisados, interpretados, compreendidos ou classificados. Por serem compartilhados em volumes enormes, a interpretação manual torna-se impossível. Uma das formas de analisar este grande volume de texto é agrupá-lo por tópico. Com isso, faz-se necessário a utilização de técnicas de modelagem de tópicos, conjunto de algoritmos de aprendizagem de máquina não supervisionada. Uma das abordagens mais atuais de extração de tópicos é implementada pelo algoritmo Bertopic. O Bertopic é uma técnica de modelagem de tópicos que utiliza transformadores e `c-TF-IDF` para criar grupos densos, facilitando a interpretação dos tópicos, mantendo palavras nas descrições dos tópicos. Deste modo, neste trabalho se apoiará na extração de tópicos utilizando o Bertopic. A coleção de dados de entrada serão *tweets* sobre o Haiti. O objetivo é fazer uma análise exploratória sobre os assuntos discutidos sobre o país no Twitter.

Palavras-chave: Modelagem de tópicos. Twitter. Haiti. Bertopic.

ABSTRACT

With the growth of internet access of the world population, large volumes of data, mainly in the format of texts, have been shared daily. Be it through social networks, blogs or forums. This data can be analyzed, interpreted, understood or classified. Because they are shared in huge volumes, manual interpretation becomes impossible. One of the ways to analyze this large volume of text is to group it by topic. This requires the use of topic modeling techniques, a set of unsupervised machine learning algorithms. One of the most current approaches to topic extraction is implemented by the Bertopic algorithm. Bertopic is a topic modeling technique that uses transformers and `c-TF-IDF` to create dense groups, facilitating the interpretation of topics, keeping words in the descriptions of topics. Thus, this work will be based on the extraction of topics using Bertopic. The collection of input data will be tweets about Haiti. The goal is to make an exploratory analysis on the topics discussed about the country on Twitter.

Keywords: Topic modeling. Twitter. Haiti. Bertopic.

RÉSUMÉ

Avec la croissance de l'accès à Internet de la population mondiale, de grands volumes de données, principalement dans le format des textes, ont été partagés quotidiennement. Que ce soit via des réseaux sociaux, des blogs ou des forums. Ces données peuvent être analysées, interprétées, comprises ou classifiées. Comme ils sont partagés en volumes énormes, l'interprétation manuelle devient impossible. Une des façons d'analyser ce grand volume de texte est de le regrouper par sujet. Il est donc nécessaire d'utiliser des techniques modèle de thèmes, un ensemble d'algorithmes d'apprentissage de machine non supervisés. L'une des approches les plus actuelles de l'extraction de sujets est mise en œuvre par l'algorithme Bertopic. Bertopic est une technique de modélisation des sujets qui utilise des transformateurs et `c-TF-IDF` pour créer des groupes denses, facilitant l'interprétation des sujets, gardant des mots dans les descriptions des sujets. Ainsi, ce travail s'appuiera sur l'extraction de sujets utilisant le Bertopic. La collection de données d'entrée sera *tweets* sur Haïti. L'objectif est de faire une analyse exploratoire des sujets discutés sur le pays sur Twitter.

Mots-clé: Modèles de thèmes. Twitter. Haïti. Bertopic.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

1.1 TEMA

O tema deste trabalho é a utilização do algoritmo *Bertopic* como técnica de modelagem de tópicos para identificar os principais assuntos discutidos sobre o Haiti nas redes sociais, particularmente no Twitter em língua portuguesa.

1.2 PROBLEMATIZAÇÃO

Atualmente, dados estão sendo compartilhados na web entre usuários que não precisam estar necessariamente na mesma posição geográfica. Esses dados podem transmitir informações referentes a diversos assuntos mundiais como, por exemplo, politica, jogos e cultura. Geralmente, os dados circulam na web no formato de textos em redes sociais, blogs e fóruns e são escritos de maneira livre. O Twitter é um exemplo de rede social muito utilizada pelos usuários a fim de disseminar conteúdos na internet a respeito de assuntos de interesse. Como o acesso a internet está crescendo e que os dados estão sendo compartilhados constantemente, esses dados podem ser classificados, interpretados ou agrupados por tópicos semelhantes.

O agrupamento dos dados ou documentos em tópicos similares, por serem dados compartilhados em quantidade enorme, requer muito tempo para a análise manual. Assim, visando minimizar o tempo de realização dessa tarefa, são desenvolvidas técnicas de aprendizado de máquina não supervisionado. Essas técnicas são algoritmos de modelagem de tópicos que visam analisar palavras em coleção de documentos a fim de descobrir temas ou assuntos principais onde esses documentos são, geralmente, textos desestruturados (BLEI, 2012).

1.3 OBJETIVOS

1.3.1 Objetivo Geral

Explorar e analisar os principais assuntos discutidos sobre o Haiti na plataforma do Twitter usando *Bertopic* como técnica de modelagem de tópicos.

1.3.2 Objetivos Específicos

- Definir a *string* para filtrar *tweets*.
- Coletar os dados no período de datas a ser definido;
- Realizar pré-processamento dos dados;
- Codificar e treinar o modelo;

- Encontrar o melhor número de tópicos a serem extraídos;
- Extrair os tópicos com o modelo *Bertopic*;
- Rotular os tópicos extraídos; e
- Analisar e classificar os resultados.

1.4 JUSTIFICATIVA

A modelagem de tópicos vem trazendo uma facilidade de análise de documentos ou coleções de documentos pelos seus algoritmos de aprendizado de máquina não supervisionado. Esses algoritmos são desenvolvidos com a perspectiva de interpretar, analisar ou classificar palavras em grande quantidade de dados que não possuem rótulos, tarefa que seria impossível realizar por humanos sem o uso de modelagem de tópicos.

Desta forma, ao desejar descobrir padrões ou realizar classificação de volumes de textos é imprescindível usar técnicas de modelagem de tópicos, pois a tarefa manual é custosa e suscetível a erros. Essas classificações podem ser relevantes para descobrir assuntos mais abordados sobre determinados eventos, assuntos de atualidades, opiniões sobre diversos ramos da sociedade.

O Haiti é um país que tem passado por muitos eventos ou desastres ao longo dos tempos. Sejam naturais ou causados por ações humanas. Dentre os desastres naturais, o país conheceu terremotos, onde o mais devastador foi de 12 de janeiro de 2010, e várias inundações que devastaram cidades diferentes. O assassinato do presidente Jovenel Moïse no dia 07 de julho de 2021 é um dos eventos causados por ações humanas. Para tanto, é necessário descobrir o quê estão mais falando sobre o país na comunidade lusófona.

Nesse contexto, com a dispersão de dados na internet sobre os diversos acontecimentos do Haiti, o uso de uma técnica de modelagem de tópicos pode facilitar o entendimento dos assuntos mais abordados nas redes sociais em língua portuguesa, principalmente no Twitter, por meio de agrupamentos das palavras relacionadas dos documentos por tópicos.

1.5 ESTRUTURA DO TRABALHO

O restante deste trabalho é apresentado da seguinte forma: O Capítulo 2 apresenta o país da pesquisa deste trabalho e definições sobre a rede social utilizada: Twitter. O capítulo 3 apresenta os conceitos de modelagem de tópicos, de tópicos, de documentos e do algoritmo Bertopic. No capítulo 5 é apresentada a metodologia contendo as etapas para alcançar os objetivos do trabalho. Por fim, é apresentado o cronograma previsto para a realização das atividades previstas.

2 HAITI E TWITTER

Este capítulo apresenta alguns conceitos importantes para o entendimento deste trabalho. Inicialmente, é apresentado o país Haiti por ser foco dos *tweets* que serão estudados. Em seguida, é apresentada a rede social Twitter.

2.1 HAITI

A República do Haiti, comumente chamada Haiti, é um país do caribe. O país é a parte oeste da ilha Hispaniola partilhada com a República Dominicana (este último fica no lado leste). Em termos de população e de área, o Haiti é o terceiro país do caribe (depois de Cuba e da República Dominicana) com uma superfície de 27.750 quilômetros quadrados e uma população ativa estimada de cerca de 11 milhões de habitantes. A capital do Haiti é Porto-Príncipe e os idiomas oficiais são francês e crioulo (WIKIPEDIA, 2022a). A Figura 1 apresenta a localização do Haiti em relação às Américas.

Figura 1 – Posição geográfica da República do Haiti.

Fonte: (WIKIPEDIA, 2022a)

2.1.1 Breve Histórico

A ilha Hispaniola, habitada por populações indígenas, foi descoberta em 1492 por Cristovão Colombo. Essas populações foram massacradas pelo trabalho forçado (mineração de ouro) dos espanhóis. Para substituir essa força de trabalho, os colonos recorrem aos escravos africanos. Juntamente com os mulatos, os escravos africanos são os principais ancestrais da grande maioria dos haitianos.

Em meados do século XVI, os franceses se estabeleceram nas terras abandonadas pelos espanhóis apesar dos esforços concentrados desses últimos para repeli-los. Os franceses, uma vez instalados na ilha, também recorreram aos escravos africanos mas, para trabalhar nas plantações de café e açúcar. Afinal, em 1697, os espanhóis cederam e reconheceram a soberania francesa sobre a parte ocidental da ilha onde é fundada a capital, Porto Príncipe, em 1749.

Dentre todas as colonias do Novo Mundo, a chamada "la Saint-Domingue française"tornou-se a mais lucrativa, mesmo a frente dos Estados unidos.

Em 1791, os negros, liderados por Jean-Jacques Dessalines, Henri Christophe, Alexandre Pétion e Toussaint Louverture, se uniram, revoltaram e pegaram armas contra a França. O fim dessa guerra foi marcado pela capitulação do exército francês em 1803. Portanto, alguns meses depois, ou seja no dia 01 de janeiro de 1804, na cidade de Gonaïves, foi proclamada a independência, tornando o Haiti a primeira república negra livre.

Se a declaração do ato da independência, redigida em 1804 tinha como intenção "Assegurar para sempre os nativos do Haiti um governo estável", os fatos provaram o contrário. Já que, entre 1804 a 1957, foram 37 chefes de estado e dentre eles, 24 foram assassinados ou derrubados. O primeiro deles, Jean-Jacques Dessalines, líder principal da revolta contra a França, permaneceu no cargo durante somente dois (2) anos, ele foi assassinado no dia 17 de outubro de 1806. O recorde de longevidade é de Jean-Pierre Boyer que governou durante 25 anos (TWITTER, 2022a).

2.1.2 Terremoto de 2010

No dia 10 de Janeiro de 2010, às 16 horas e 53 minutos e 10 segundos (hora local do Haiti), ocorreu um terrível terremoto com magnitude de 7,0 a 7,3 que devastou o Haiti. O epicentro do terremoto está localizado a aproximadamente a 25,3 km de Porto-Príncipe, capital do país. E o seu hipocentro foi localizado a uma profundidade de 10 km.

A ilha Hispaniola encontra-se em uma zona sismicamente ativa, entre duas placas tectônicas que são respectivamente: a placa norte-americana ao norte e a placa caribenha ao sul. Tal zona possui falhas que são deslizamentos sinistrais e falhas de compressão. Assim, a causa do terremoto foi uma ruptura de uma falha.

Nesse terremoto, foram muitas mortes, vítimas, ferimentos, pessoas amputadas ou com pertubações psicológicas. Conforme os dados divulgados no dia 09 de fevereiro de 2010 pela Ministra das Comunicações, Marie-Laurence Jocelyn Lassegue, foram mais de 280 mil mortos, 300 mil feridos e 1,3 milhão de pessoas sem abrigo. Além disso, muitas instituições importantes do país foram destruídas tais: o palácio nacional onde o presidente exerce o seu trabalho e a Catedral de Notre-Dame de Porto-Príncipe (WIKIPEDIA, 2022b).

2.1.3 Assassinato Presidente 2021

Jovenel Moïse, nascido no dia 26 de 1968 na comuna Trou du Nord / Nordeste do Haiti, presidente da República do Haiti, eleito em 2016 e tomou posse no dia 07 de fevereiro de 2017, foi assassinado na madrugada entre o dia 06 e 07 de julho de 2021 (por volta de uma hora em sua residência privada a Pétionville) por um grupo armado não identificado onde a primeira dama foi ferida conforme a declaração do chefe do governo haitiano (TROUILLARD, 2022).

Horas depois, esta morte deixou o país em desordem. "Não sabemos quem está dirigindo", diz um membro de um partido oposto ao governo do Moise. Com essa drama, o país afundou novamente em uma crise após vários meses de crises institucionais pendentes. O parlamento não era mais funcional há um bom tempo for falta de organização de eleições legislativas a tempo. A tragédia deixou todo mundo preocupado pois, ninguém sabe quem está no comando, quem controla o quê. Em pouco tempo após isso, a República Dominicana ordenou o fechamento imediato de sua fronteira com a República do Haiti (BIASSETE, 2022).

2.2 TWITTER

Em 2006, segundo o site oficial (TWITTER, 2022b), foi lançado o Twitter, uma das redes sociais mais populares da atualidade. Conhecido também como um microblog, o Twitter permite que usuários compartilhem perguntas, sentimentos, ideias e opiniões referentes a vários assuntos como política, saúde, economia, sociedade, música, cinema, celebridades e jogos por meio de *tweets*. Um *tweet* é uma publicação realizada na plataforma que, geralmente, compõe-se de textos, fotos, links ou vídeos

Os *tweets* possuem um limite máximo de 280 caracteres e são visíveis pelo próprio autor, seguidores ou até mesmo por pessoas interessadas no determinado assunto. Para destacar melhor o assunto principal do *tweet*, geralmente, usam-se as *hashtags* (#), caracteres que contribuem na exibição de *tweets* ao realizar-se a busca de assuntos ou temas populares. Atualmente, mais de 500 milhões de *tweets* são publicados diariamente em mais de quarenta (40) idiomas. Isto se deve ao crescimento da população ativa na plataforma com uma estimativa de 192 milhões de usuários ativos diariamente em 2020.

Todo usuário da plataforma pode ter *followers* e *following*. Chama-se de *followers*, usuários que seguem o usuário em questão , ou seja, usuários que se inscrevem no perfil de outros usuários para que seja possível visualizar novos conteúdos publicados pelos usuários. Já, *following* são os usuários que o usuário do perfil está seguindo para visualizar ou interagir em conteúdos onde ele não é o usuário principal.

A Figura 2 mostra o exemplo de um perfil no microblog com o destaque dos itens mais relevantes. Na parte a), é mostrado o nome oficial inserido pelo usuário. No item b), é exibida a quantidade total de *tweets* publicados pelo usuário na plataforma. O item c) apresenta a foto do perfil. O item d) mostra o nome do usuário formado por um nome escolhido pelo usuário seguido pelo caractere @. Esse nome de usuário deve ser único na plataforma. No item e), são indicados os dados da biografia do usuário. Por fim, nos itens f) e g) respectivamente, é denotada a quantidade de *following* e *followers*.

Na Figura 3, é mostrado o exemplo de um *tweet* publicado sobre o Haiti por um ex-comandante das Nações Unidas para a estabilização do Haiti. Na imagem em questão, são destacadas as principais partes de um *tweet*: no item a), de esquerda a direita, são mostrados, sequencialmente, o nome oficial do perfil, o nome do usuário e a data de publicação do *tweet*;

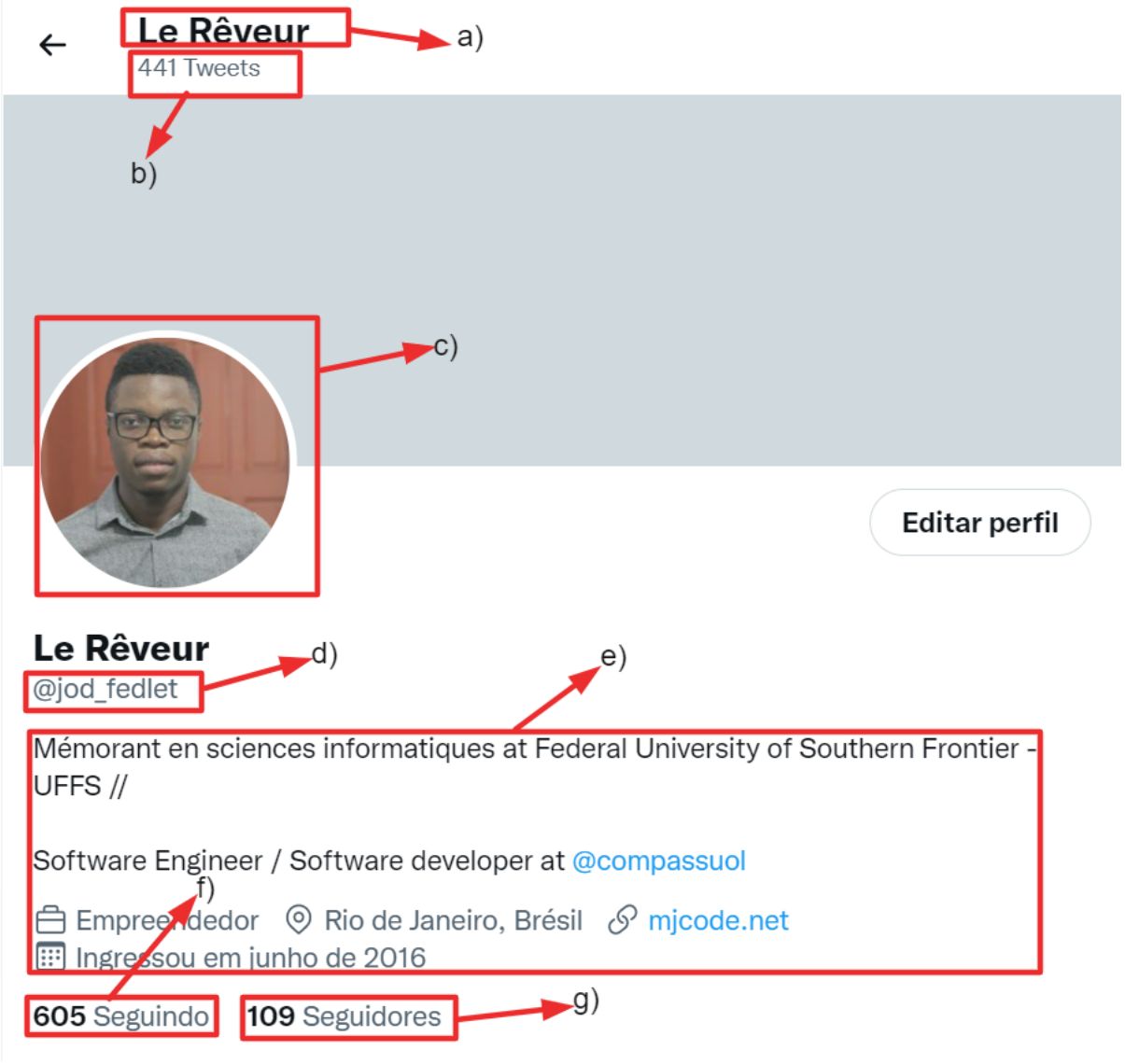


Figura 2 – Exemplo de perfil no twitter.

no item b), é exibido o conteúdo oficial do *tweet* com o tema principal em negrito e por fim, no item c), de esquerda a direita, são mostrados os dados de interações de outros usuários ao *tweet* publicado que são, respectivamente, quantidade de comentários (28), quantidade de vez que foi retweetado (80), número de curtidas (523) e número de compartilhamento (0) do *tweet*.

2.3 CONSIDERAÇÕES FINAIS

Neste capítulo, foram apresentados os principais conceitos sobre o Haiti e o Twitter para facilitar o entendimento do trabalho. Tendo o Haiti como embasamento principal dos temas dos tópicos a serem analisados, é fundamental apresentá-lo junto com alguns fatos importantes ocorridos no país ao longo dos tempos. Sendo assim, foi apresentado um breve histórico sobre o país durante o período pré-colonial onde a parte ocidental da ilha Hispaniola foi habitada por populações indígenas; o período colonial que, inicialmente, foi marcado pela presença dos

Figura 3 – Exemplo de *tweet* publicado sobre o Haiti.

espanhóis na ilha mas, no final a parte ocidental foi tomada pela França e a revolta dos escravos negros vindos da África liderada por, especialmente, Jean-Jacques Dessalines; o período pós-colonial marcado pela declaração do ato da independência da parte ocidental que atualmente é o Haiti e pela instabilidade do país após a independência. Também, foram apresentados o terremoto que ocorreu no dia 12 de janeiro de 2010 onde mais de 200 mil pessoas faleceram e o assassinato recente do presidente da República do Haiti, sua excelência Jovenel Moïse, no dia 07 de Julho de 2021.

O Twitter, por ser a base onde os documentos serão coletados, foi necessário apresentá-lo com as suas características. Sendo um microblog, ele serve para disseminar ideais, pensamentos no formato de textos, imagem, vídeo por meio de publicação a respeito um determinado assunto. Essas publicações se chamam *tweet* com um limite de até 280 caracteres. Com essa limitação de caracteres, os *tweets* são considerados como documentos curtos. Portanto, neste trabalho, será utilizada a base de dados do Twitter para identificar os principais tópicos discutidos sobre o Haiti.

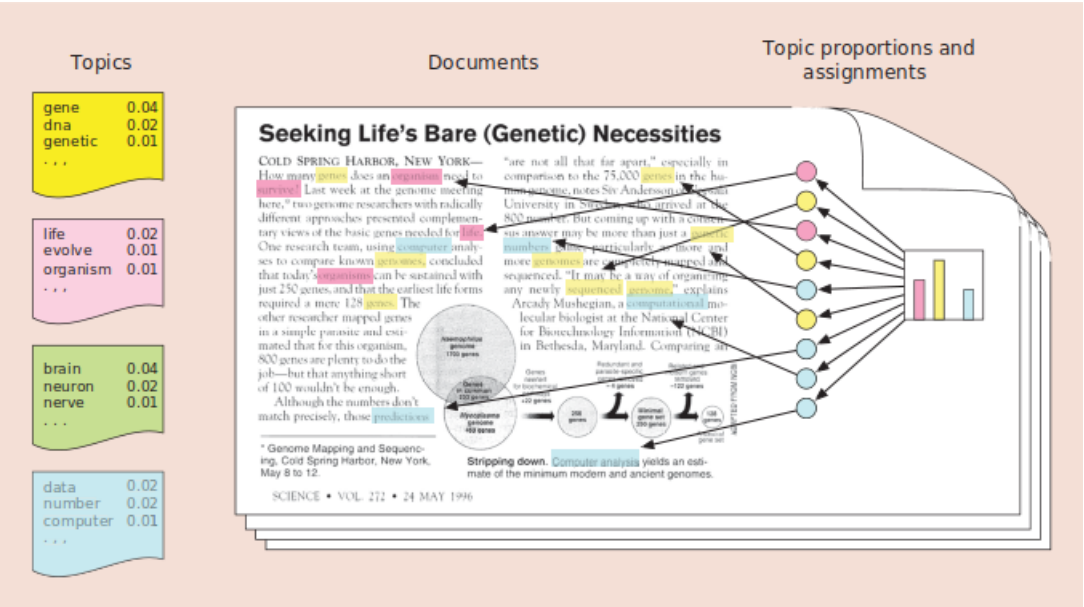
3 MODELAGEM DE TÓPICOS

A modelagem de tópicos é um conjunto de algoritmos de aprendizagem de máquina não supervisionada conhecidos como modelos que são utilizados para analisar, agrupar, por semelhanças, dados em grandes volumes de textos ou coleção de documentos. Estes modelos se baseiam na ideia básica de que um documento é uma mistura de tópicos mas, com hipóteses levemente diferentes (STEYVERS; GRIFFITHS, 2007). Os documentos, por não possuírem rótulos, são bons dados de aplicação dos algoritmos de modelagem de tópicos que, inicialmente, foram desenvolvidos com essa perspectiva.

Algoritmos de modelagem de tópicos são métodos estatísticos que analisam as palavras oriundas de um documento ou coleção de documentos para descobrir os assuntos ou temas principais. Os algoritmos podem ser aplicados em grandes coleções de documentos desestruturados. Além dos documentos textuais, podem-se aplicar estes algoritmos em outros tipos de dados como imagem para encontrar padrões latentes (ocultos) (BLEI, 2012).

Ainda, segundo (BLEI, 2012), os algoritmos de modelagem de tópicos não requerem nenhum tipo de documento rotulado. Baseados nos documentos originais, eles conseguem agrupar os documentos por tópicos utilizando distribuição estatística da aparição das palavras nos documentos. Os algoritmos de modelagem de tópicos ajudam a analisar e organizar grandes volumes de dados, tarefa que seria impossível realizar por seres humanos.

Figura 4 – Exemplo de modelagem de tópicos.



Fonte: (BLEI, 2012)

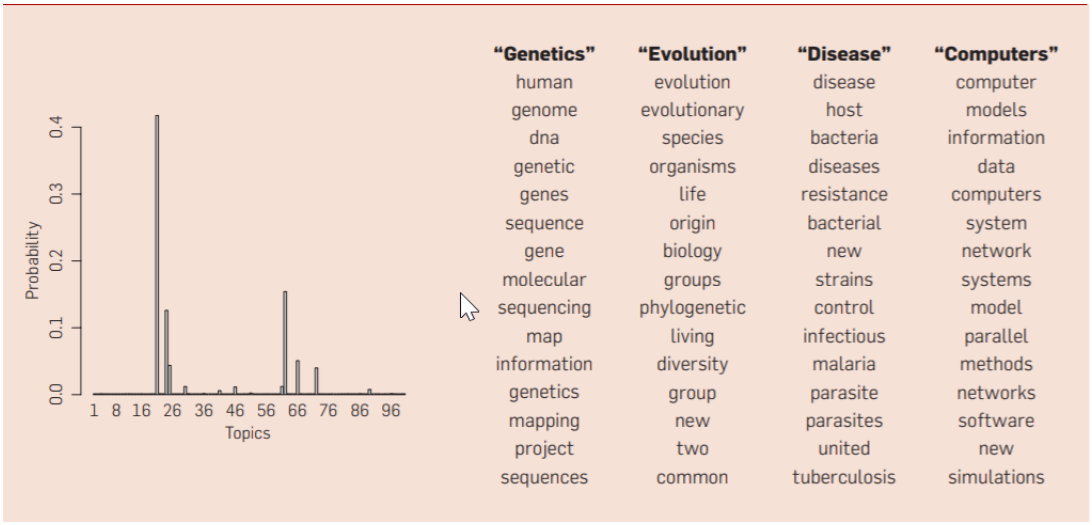
Segundo (BLEI, 2012), a Figura 4 apresenta um artigo intitulado *Seeking Life's Bare (Genetic) Necessities* que fala sobre o uso da análise de dados para determinar o número de genes que um organismo precisa pra sobreviver, em um sentido evolucionário. Foram

ressaltadas, manualmente, palavras diferentes que são usadas no artigo. Dentre elas, palavras sobre análise de dados, como *computador*, *previsão* estão em azul; em rosa, palavras sobre evolução biológica, como *vida* e *organismo*; em amarelo, são destacadas as palavras referentes a genética, como *sequenciado* e *genes*. Se fosse destacar cada palavra do artigo, daria pra perceber uma combinação de genética, análise de dados e biologia evolucionária em diferentes proporções. Além disso, sabendo que o artigo combina tais tópicos ajudaria a situar em uma coleção de artigos científicos. Para identificar os tópicos citados acima, foi utilizado o modelo Latent Dirichlet Allocation (LDA) que é considerado como o modelo de tópico mais simples. A ideia dele é que documentos exibem vários tópicos. A seguir, serão apresentados os conceitos de tópicos, documentos e apresenta uma abordagem de extração chamada *Bertopic*.

3.1 TÓPICOS

Segundo (BLEI, 2012), os tópicos são uma distribuição de palavras acerca de um documento. Tendo como exemplo, tópicos sobre genética são formados de palavras referentes à genética com alta probabilidade sobre genética e tópicos sobre biologia evolucionária são formados de palavras relacionadas à biologia evolucionária. A Figura 5 apresenta o provável rótulo de quatro tópicos, em negrito, de um documento que são respectivamente *Genética*, *Evolução*, *Doença* e *Computador*. Embaixo de cada um desses tópicos, são apresentadas as 15 palavras mais frequentes, ou seja palavras com mais probabilidade estatística de aparição, em ordem decrescente.

Figura 5 – Exemplo de tópicos.

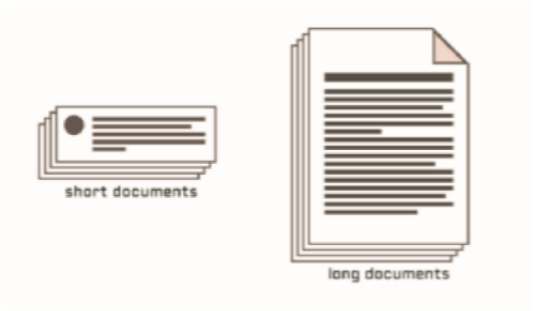


Fonte: (BLEI, 2012)

3.2 DOCUMENTOS

De acordo com (DAVID M. BLEI ANDREW Y. NG, 2003), documentos são uma mistura aleatória sobre tópicos, ou seja, é possível extrair vários tópicos a partir de um mesmo documento. Um documento pode ser curto ou longo dependendo da sua fonte e, geralmente, é composto de texto. Como neste trabalho a base onde os documentos serão extraídos para identificar tópicos é o *Twitter*, serão usados documentos curtos pelo fato de que os *tweets* possuem o limite máximo de até 280 caracteres.

Figura 6 – Exemplo de documentos



Fonte: (PEREIRA, 2019)

Na Figura 6 é mostrado o exemplo de documentos curto e longo da esquerda para a direita, respectivamente. O documento curto pode ser representado por *tweets*, postagens, comentários, enquanto artigos científicos, notícias são exemplos de documentos longos (PEREIRA, 2019).

Neste trabalho, *Bertopic* é o modelo a ser utilizado para identificar tópicos discutidos sobre o Haiti em documentos curtos usando a base de dados do *Twitter*.

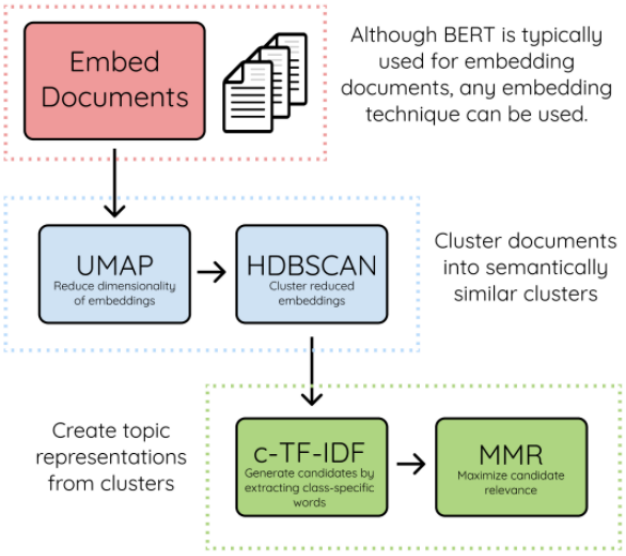
3.3 MODELAGEM DE TÓPICOS COM BERT (BERTOPIC)

O *BERT* (Bidirectional Encoder Representations from Transformers) é uma arquitetura neural profunda que possui como característica fundamental o fato de ser bidirecional, o que facilita a compreensão de uma palavra ou do texto em geral, analisando as palavras adjacentes em ambas as direções. Por ser bidirecional, permite um bom desempenho mesmo com uma pequena amostra do texto diferentemente dos modelos unidirecionais que demandam uma amostra maior para ter resultados satisfatórios por não conseguirem analisar simultaneamente palavras da esquerda e da direita. O transformador, unidade básica do *BERT*, possui dois mecanismos principais: um codificador que faz a leitura do texto de entrada; e um decodificador, que utiliza um modelo preditivo para a compreensão do texto (CAPELLARO, 2021).

Segundo (GROOTENDORST, 2020), *Bertopic* é uma técnica de modelagem de tópicos, baseada no *BERT*, que utiliza transformadores e *c-TF-IDF (term frequency - inverse document frequency)* para criar grupos ou clusters densos permitindo tópicos facilmente interpretáveis,

mantendo palavras importantes na descrição de tópicos. Na Figura 7, pode-se observar as três etapas do algoritmo *BERTopic*.

Figura 7 – Fluxo de funcionamento do algoritmo *BERTopic*.



Fonte: (GROOTENDORST, 2020)

As três etapas da Figura 7 podem ser descritas como:

- Em rosa, é mostrada a primeira etapa que é a extração de incorporações de documentos com *BERT* ou com qualquer outra técnica de incorporação;
- Em azul, é apresentada a segunda etapa que está subdividida em duas fases:
  1. *UMAP* (redução da dimensionalidade das incorporações geradas na etapa 1);
  2. *HBDSCAN* (agrupamento, clusterização das incorporações reduzidas).
- Em verde, é destacada a terceira e última etapa do algoritmo do *BERTopic* que é a criação dos tópicos, onde cada cluster (grupo) é convertido em um único documento em vez de um conjunto de documentos. Perceba que após a clusterização, a técnica TF-IDF é aplicada em cada *cluster* independentemente, utilizando a forma c-TF-IDF, ou seja, a importância das palavras é calculada por *cluster* (GROOTENDORST, 2020).

Resumidamente, BERTopic é uma técnica de modelagem de tópicos que utiliza transformadores e `c-TF-IDF` para criar agrupamentos densos, permitindo tópicos interpretáveis, mantendo palavras importantes nas descrições dos tópicos (GROOTENDORST, 2020).

3.4 CONSIDERAÇÕES FINAIS

Os *tweets* são textos escritos livremente e, portanto, desestruturados. Assim, são próprios para serem aplicados em modelagem de tópicos. Posto isso, neste trabalho será utilizada a

abordagem *Bertopic* que é uma técnica de extração de tópicos em coleção de documentos. A coleção de documentos (*tweets*) utilizada como entrada será obtida na base de dados do Twitter sobre o Haiti, ou seja, *hashtags* que indiquem que um determinado *tweet* é sobre esse país.

4 TRABALHOS RELACIONADOS

Neste capítulo, serão apresentados os quatro trabalhos relacionados ao tema que foram selecionados a fim de direcionar o estudo.

No trabalho de (PEREIRA, 2019), é realizada uma análise exploratória de *tweets* extraídos na base de dados do *Twitter*, utilizando o modelo *BTM (Biterm Topic Model)*, tendo como objetivo geral descobrir os tópicos discutidos durante o período dos Jogos Olímpicos de 2016. Alguns dos objetivos específicos do trabalho são:

- definir o método a ser utilizado;
- definir o período de datas para a obtenção dos *tweets*;
- realizar o pré-processamento dos dados;
- verificar e analisar os principais assuntos referentes aos Jogos Olímpicos Rio 2016.

Para extrair os *tweets* na base do *Twitter*, foram utilizadas as seguintes strings de busca: "rio2016", "olimpiadas", "olimpiada", "ceremoniadeabertura", "ceremoniadeencerramento", "jogosolimpicos", "olympics", "olympicgames", "openingceremony", "closingceremony". Foi definido também, o período de dias de 02 de agosto de 2016 a 24 de agosto de 2016 para realizar as devidas buscas. Como resultado, foi obtido um total de 2.806.785 *tweets* divididos em 230 arquivos, ou seja, um arquivo para cada palavra da *string* de busca (10) multiplicado pela quantidade de dias (23). Em seguida, os arquivos foram divididos em grupos de 9, um para cada intervalo definido (a Figura 8 apresenta os 9 intervalos).

Figura 8 – Exemplo de *tweets* extraídos na base de dados do Twitter.

Arquivo	Dias	Tamanho	Quantidade de Tweets
1	02.08.2016 a 04.08.2016	27,2MB	128.527
2	05.08.2016	65,9MB	334.125
3	06.08.2016 a 08.08.2016	127MB	624.444
4	09.08.2016 a 11.08.2016	51MB	239.912
5	12.08.2016 a 14.08.2016	52,1MB	245.149
6	15.08.2016 a 17.08.2016	64,7MB	311.581
7	18.08.2016 a 20.08.2016	59,2MB	279.432
8	21.08.2016	53,9MB	260.063
9	22.08.2016 a 24.08.2016	74MB	350.763

Fonte: (PEREIRA, 2019)

A Figura 8 apresenta o conjunto de *tweets* obtidos após a realização do pré-processamento dos dados onde os *tweets* extraídos na base do Twitter são divididos em 9 arquivos e agrupados em fatia de dias.

Tendo os *tweets* extraídos, divididos em arquivos e agrupados por intervalo de dias, foi realizado o pré-processamento dos dados, processo em que é realizada a limpeza dos dados. Para

isso, foi necessário remover e/ou tratar alguns dados julgados desnecessários na continuação da análise exploratória. Os tratamentos realizados são os seguintes:

- Conversão dos *tweets* em letras minúsculas;
- Remoção de *stop words*, pontuação e acentuação, e das URLs;
- Remoção de *tweets* contendo menos de 3 palavras, exceto quando estivesse representando país, por exemplo: br, fr, us e de *tweets* duplicados.
- Remoção de menções (palavras seguidas do caractere '@') e de *hashtags* (palavras seguidas do caractere '#');
- Remoção de dígitos com exceção de 2016 quando seguido da palavra rio.

Após o pré-processamento dos *tweets*, reduziu-se o conjunto de 2.806.785 *tweets* para 1.548.759, com um total de 8.611.136 palavras e com uma média de 5,56 palavras por *tweet*. Posteriormente ao pré-processamento dos dados, foram aplicado o algoritmo de modelagem de tópicos *BTM* para identificar os tópicos na coleção de documentos e realizado o pós-processamento dos dados onde foi definida a quantidade de tópicos que será utilizada no experimento, sendo K igual a 5, 10, 15, 20 e 30.

Na Figura 9 é mostrado o exemplo de um *tweet* obtido após a extração (lado esquerdo) contendo URL, caracteres especias como '#' e '@' e o mesmo após a aplicação da etapa de pré-processamento dos dados (lado direito).

Figura 9 – Exemplo de *tweets* antes e depois do pré-processamento dos dados.

Original	**Após limpeza**
Dia de futebol feminino nessas Olimpíadas. #Rio2016 @Arena Corinthians - Itaquerão https://www.instagram.com/p/BIyVNoOBHvD/	futebol feminino olimpiadas corinthians itaquerao

Fonte: (PEREIRA, 2019)

Como resultados do trabalho, cada arquivo de *tweets* da Figura 8 é representado por uma tabela contendo os tópicos e as suas probabilidades de ocorrência. Em cada tabela, os tópicos estão em ordem crescente, ou seja, o tópico mais provável aparece na primeira posição, seguidos da sua probabilidade de ocorrência e o conjunto de palavras relacionado ao determinado tópico. Depois disso, foi realizada a rotulação dos tópicos manualmente baseada em fatos empíricos.

A Figura 10 é uma representação do arquivo 1 da Figura 8 dos tópicos extraídos do dia 05 de agosto de 2016. É possível notar que muitas das palavras se referem a momentos ocorridos no evento, como exemplo, palavras como 'anita', 'caetano', 'gil', 'cantando', 'cantar', 'mc', 'fernanda, 'karol', 'montenegro', 'gilberto', mencionam as seguintes apresentações: Anita,

Caetano Veloso e Gilberto Gil cantando músicas de Ary Barroso e João Gilberto, Karol Conka e Mc Soffia cantando rap e Fernanda Montenegro declamando trechos de poema "A Flor e a Náusea", de Carlos Drummons de Andrade. Por outro lado, as palavras como 'pira', 'olimpica', 'acender', 'vanderlei', 'tocha', 'guga', 'cordeiro', 'lima' e 'chama', falam do momento em que a chama olímpica aparece no Maracanã nas mãos do ex-tenista Guga, e em seguida, o ex-maratonista Vanderlei Cordeiro de Lima a recebe, acendendo a pira e oficializando o início dos Jogos Olímpicos Rio 2016 (PEREIRA, 2019).

Figura 10 – Tópicos obtidos a partir dos *tweets* extraídos do dia 05 de Agosto de 2016

Rótulo	**Palavras**
Cerimônia de Abertura	pais, lindo, copa, mundo, brasileiro, povo, dinheiro, bonito, mal, festa, deus, gisele, demais, cerimonia, orgulho, maravilhosa, amo, parabens, pais, melhor, hoje, momento, atleta, anitta, caetano, gil, cantando, cantar, mc, fernanda, karol, montenegro, gilberto, jogo, medalha, ouro, futebol, esporte, selecao, olimpico, delegacao, bandeira, alemanha, mulher, vem, porta, grecia, aquecimento, indios, portugueses, global, parte, mostrar, historia, hino, musica, paulinho, viola, nacional, zeca, flamengo, pira, olimpica, acender, vanderlei, tocha, guga, cordeiro, pele, lima, chama, maracana, janeiro, cidade, estado, paises, samba, aula, carnaval, geografia, escola, falar, regina, ingles, case, falando, portugues, tremendo, santos, fala, homem.
Mídia	galvao, boca, cala, gloria, maria, globo, ouvir, bueno, falando, falar.
Político	temer, vaia, dilma, presidente, vaiado, golpista, michel, lula, medo, povo.
Transmissão das Olimpíadas	pokemon, vendo, hoje, cerimonia, mundo, casa, assistir, assistindo, estar, tv.

Fonte: (PEREIRA, 2019)

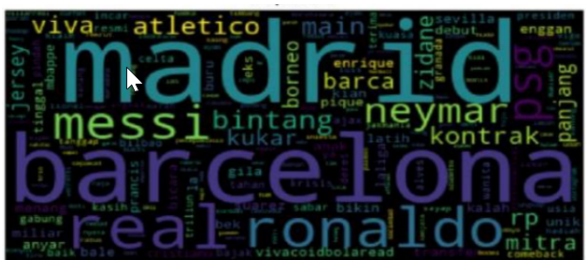
Em (HIDAYATULLAH et al., 2018), foi utilizada a modelagem de tópicos para determinar tópicos oriundos de *tweets* das competições de futebol escritos em Bahasa (idioma oficial da Indonésia). Foram escolhidas 5 ligas: La Liga nacional da Indonésia, a Premier League da Inglaterra, a Bundesliga da Alemanha, a La Liga da Espanha e a Serie A da Itália. Os dados do estudo foram coletados em várias contas oficiais do Twitter indonésio que, em todo o tempo, atualizam sobre o futebol. Para definir o tipo dos tópicos, foi aplicado o Latent Dirichlet Allocation (LDA) como método de modelagem de tópicos.

Para obter os *tweets* na base de dados do Twitter sobre as competições supracitadas, usaram-se cinco contas oficias que são *@VIVAbola*, *@panditfootball*, *@detiksport*, *@Bolanet* e *@GOAL_ID* e definiu-se o o intervalo de tempo entre o dia 01 de janeiro de 2017 a 24 de dezembro de 2017. Com isso, foi obtido um total de 120.639 *tweets*. Após a obtenção dos tweets, foi realizado o pre-processamentos dos dados para ajustar os *tweets* que não estavam de acordo. Por fim, foi aplicado o modelo LDA.

Como resultados, usando dez como a quantidade de tópicos a serem extraídos, foram identificados tópicos relacionados a Premier League, a Bundesliga, a La Liga da Espanha, a Serie A e a Liga nacional de futebol da Indonésia conforme esperado. A Figura 11 exibe uma nuvem de palavras de um dos dez tópicos do experimento. Refere-se a rivalidade do El Clasico entre Real Madrid e FC Barcelona, os dois maiores times de futebol masculino da Espanha e dois dos melhores do mundo. Em destaques, as palavras com tamanho maior referem-se aos

dois times que são *madrid* e *barcelona* e as palavras com tamanho médio como *messi*, *ronaldo* e *neymar* referem-se aos jogadores mais famosos do El Clasico.

Figura 11 – Nuvem de palavras obtidas em um dos dez tópicos.



Fonte: (HIDAYATULLAH et al., 2018)

(ASGHARI; SIERRA-SOSA; ELMAGHRABY, 2018) propuseram um sistema adaptativo projetado usando 5 camadas e contém uma associação de algoritmos não supervisionados e supervisionados para rastrear as tendências sobre saúde nas mídias sociais, com foco nos dados de cada estado dos Estados Unidos. A rede social utilizada na qual os dados do estudo foram coletados foi o *Twitter*. LDA e Word2Vect são os métodos de modelagem de tópicos utilizados que são, respectivamente, não supervisionado e supervisionado.

Os *tweets* relacionados à saúde foram coletados por um mês totalizando 144.922. Para tal coleta, foram empregadas as palavras-chave como: *healthcare*, *health*, *doctors*, *homecare*, *digitalhealth* and *digital health*. Durante a coleta, observou-se também a presença de palavras não relacionadas à saúde. Para removê-las, foi aplicada filtragem e os *tweets* foram reduzidos a 37.910. Tendo todos os *tweets* relacionados à saúde, foram aplicadas as seguintes etapas:

- Lematização, processo no qual é realizada uma análise morfológica das palavras, por exemplo palavras da terceira posição são alteradas para a primeira e verbos no passado e futuro transformados no presente;

- *Stemming* em que as terminações flexionais, prefixo e sufixo das palavras são removidos.

Ao aplicar o LDA, constatou-se que cinco foi o melhor número de tópicos a serem extraídos.

Como resultados, cada tópico obtido possui porcentagens de interesse maior em determinados estados e menor em outros. Por exemplo, para o tópico *Care*, Califórnia e Nova York possuem uma porcentagem de 80% de interesse, ou seja são os estados mais interessados e Virgínia com 41% é o menos interessado.

O trabalho de (OH; KWON; RAO, 2010) propõe uma exploração a dinâmica de trabalho de fábrica de boatos (*fake news*) analisando dados do Twitter tendo como foco os *tweets* postados durante o terremoto no Haiti em 2010. Foram realizados processo de identificação dos tipos de conteúdo das postagens dos *tweets* por meio de codificação e processo de análises quantitativas e qualitativas. Como esquema de codificação, foi usado o Rumor Interaction Analysis Systems

(RIAS), ANOVA de Friedman como método estatístico não paramétrico para a análise quantitativa na qual é estudada a mudança de padrões de modo de comunicação e o Automap, software de visualização usado na análise de rede semântica que é uma análise qualitativa.

Para coletar os dados relacionados ao terremoto de 2010 no Haiti no microblog Twitter, usou-se *#haitiearthquake* como palava-chave e definiu o intervalo de dez (10) dias, partindo do primeiro dia do terremoto, ou seja do dia 12 de janeiro de 2010 até o dia 21 de janeiro de 2010. Após essas definições, obteve-se um total de mais de 3.000 *tweets* por dia contendo *tweets* não relacionados ao assunto principal e *tweets* postados em idiomas francês e crioulo que não são relevantes para a pesquisa. Portanto, os mesmos foram excluídos e ficou com um total de 960 *tweets* relacionados ao tema em língua inglesa.

Posteriormente, foram realizadas limpeza de dados e análises quantitativa, definindo o modelo de quatro estágio como o mais adequado.

Consequentemente, ao executar a rede semântica com os quatro conjuntos de dados igualmente divididos identificados precedentemente para explorar padrões de mudança no modo de comunicação, partindo de amostras de cada estágio, as palavras como *blog*, *CNN*, *picture*, *list*, *info*, *report* e *update* são relacionadas a declarações de autenticação e são um exemplo dos resultados obtidos.

Este trabalho, como os citados nesse capítulo, utilizará *tweets* para gerar uma coleção de documentos. Conforme já apresentado, os *tweets* serão relacionados ao país Haiti (como o trabalho de (OH; KWON; RAO, 2010)) porém utilizando a abordagem *BERTopic* para a descoberta dos tópicos.

5 METODOLOGIA

Neste trabalho, será realizada uma análise exploratória de *tweets* usando o Bertopic como técnica de modelagem de tópicos. Assim, os dados serão coletados na base Twitter e serão baseados nas postagens realizadas exclusivamente sobre o Haiti. A seguir, estão listados os passos necessários para a realização do dito estudo.

5.1 INÍCIO DOS ESTUDOS

Inicialmente, o orientando contatou o orientador para mostrar a sua intenção de iniciar os seus estudos na grande área de aprendizagem de máquina e que ele fosse o seu guia ao longo do processo. Logo, o pedido foi discutido junto ao orientando e aceito pelo orientador. Assim, os estudos foram iniciados.

5.1.1 Definição do tema

Ao iniciar os estudos, o orientador foi quem sugeriu o tema no qual este trabalho está se baseando por meio de discussão com o orientando. Assim, o tema foi definido e surgiu-se a necessidade de iniciar as pesquisas dos trabalhos científicos para fundamentar o trabalho.

5.1.2 Pesquisa bibliográfica

Tendo o tema definido, iniciou-se a pesquisa bibliográfica. Previamente, alguns trabalhos como (DAVID M. BLEI ANDREW Y. NG, 2003), (BLEI, 2012) e (STEYVERS; GRIFFITHS, 2007) foram escolhidos pelo orientador a fim de entender melhor o tema de pesquisa. Em seguida, foi usada a plataforma de busca *Google Scholar* [1] para buscar trabalhos relacionados usando as seguintes strings de busca "Twitter e modelagem de tópicos"e "Twitter and topic modeling", respectivamente em português e inglês, também propostas pelo orientador. Dos resultados da string do idioma português, foi escolhido o trabalho realizado por (PEREIRA, 2019) e da string do idioma inglês, escolheram-se os trabalhos de (HIDAYATULLAH et al., 2018) e (ASGHARI; SIERRA-SOSA; ELMAGHRABY, 2018) mediante aprovação do orientador. Por fim, o trabalho de (OH; KWON; RAO, 2010) foi proposto pelo orientador e foi considerado primordial por tratar *tweets* relacionados a eventos ocorridos no Haiti. Consequentemente, foram quatro trabalhos principais relacionados que serviram como base para direcionar os estudos.

[1] https://scholar.google.com/

5.2 COLETA DOS DADOS

Etapa na qual serão coletados os *tweets* na base de dados de Twitter por meio de ferramenta disponibilizada pela própria plataforma. Desta forma será usada a API (Aplication Programming Interface), ferramenta de uso externo [2], buscando *tweets* a partir da palavra-chave: '#Haiti' em língua portuguesa com indicação de período de datas para a realização da coleta.

5.3 PRÉ-PROCESSAMENTO DOS DADOS

Etapa que visa tratar os dados obtidos na fase da coleta. Sendo assim, os *tweets* sofrerão alteração seja pelo processo de lematização, stemming ou por outra técnica como remoção de *tweets* ou parte de *tweets* considerados como dados indesejados ou desnecessários para as próximas etapas.

5.4 APLICAÇÃO DO MODELO BERTOPIC

Nesta etapa, primeiramente será realizada a codificação do modelo Bertopic e depois treiná-lo com as coleções de documentos de textos obtidos e pré-processados nas fases anteriores.

5.5 EXPERIMENTAÇÃO

Esta etapa consiste em identificar a quantidade de tópicos a serem utilizados, testar o modelo codificado com as coleções. A fase de identificação do número de tópicos é muito relevante pelo fato de que uma quantidade muito pequena ou muito grande as vezes não representa melhor as coleções. Portanto, é muito importante definir uma quantidade adequada.

A fim de identificar os tópicos, testes do modelo *Bertopic* serão aplicados com as coleções indicando o número de tópicos definido anteriormente. Assim, o modelo retornará os agrupamentos contendo as palavras que melhor representam os tópicos com as suas probabilidades. Por fim, será realizado processo de rotulação dos tópicos baseado em fatos empíricos a partir das palavras de cada agrupamentos.

5.6 ANÁLISE DOS RESULTADOS

Ao ter os tópicos identificados pelo modelo, será realizada a análise dos resultados, outra etapa que permitirá alcançar os objetivos deste trabalho. Nessa análise, serão explicados, interpretados ou comparados os tópicos obtidos a fim de situá-los com os acontecimentos, fatos ou eventos reais ocorridos no Haiti pois, o modelo é incapaz de realizar tal análise.

[2] https://developer.twitter.com/en

6 CRONOGRAMA

Neste capítulo, é apresentado o cronograma previsto para a realização das etapas definidas acima e está estruturado conforme a tabela abaixo.

Atividades	Nov		Dez		Jan		Fev		Mar		Abr		Mai		Jun		Jul		Ago		Set	
Início dos estudos	X	X	X	X	X	X	X	X	X	X												
Coleta dos dados											X	X										
Pré-processamento dos dados													X	X								
Aplicação do Bertopic														X	X							
Experimentação															X	X	X					
Análise dos resultados																	X	X	X	X		
Redação da monografia												X	X	X	X	X	X	X	X	X	X	X

Tabela 1 – Cronograma de realização das atividades propostas

REFERÊNCIAS

ASGHARI, Mohsen; SIERRA-SOSA, Daniel; ELMAGHRABY, Adel. Trends on health in social media: Analysis using twitter topic modeling. In: IEEE. 2018 IEEE international symposium on signal processing and information technology (ISSPIT). [S.l.: s.n.], 2018. p. 558–563.

BIASSETE, Gilles. **Assasinato do presidente Jovenel Moise. La Croix**. 2022. Disponível em: <https://bit.ly/3Ay1ybs>. Acesso em: 19 jan. 2022.

BLEI, David M. Probabilistic topic models. **Communications of the ACM**, Association for Computing Machinery (ACM), v. 55, n. 4, p. 77, abr. 2012. DOI: 10.1145/2133806.2133826. Disponível em: <https://doi.org/10.1145/2133806.2133826>.

CAPELLARO, Leonardo. Análise de polaridade e de tópicos em tweets no domınio da polıtica no Brasil. Universidade Federal de São Carlos, 2021.

DAVID M. BLEI ANDREW Y. NG, Michael I. Jordan. Latent Dirichlet Allocation, 2003.

GROOTENDORST, Maarten. **BERTopic: Leveraging BERT and c-TF-IDF to create easily interpretable topics.** [S.l.]: Zenodo, 2020. DOI: 10.5281/zenodo.4381785. Disponível em: <https://doi.org/10.5281/zenodo.4381785>. Acesso em: 25 jan. 2022.

HIDAYATULLAH, Ahmad Fathan et al. Twitter topic modeling on football news. In: IEEE. 2018 3rd International Conference on Computer and Communication Systems (ICCCS). [S.l.: s.n.], 2018. p. 467–471.

OH, Onook; KWON, Kyounghee Hazel; RAO, H Raghav. An exploration of social media in extreme events: Rumor theory and Twitter during the Haiti earthquake 2010, 2010.

PEREIRA, Mariana. Análise exploratória de tweets utilizando modelagem de tópicos para textos curtos: caso Olimpıadas Rio 2016. Universidade Federal da Fronteira Sul, 2019.

STEYVERS, Mark; GRIFFITHS, Tom. Probabilistic topic models. **Handbook of latent semantic analysis**, v. 427, n. 7, p. 424–440, 2007.

TROUILLARD, Stéphanie. **Assasinato do presidente Jovenel Moise. France24**. 2022. Disponível em: <https://bit.ly/3g49WG1>. Acesso em: 19 jan. 2022.

TWITTER. **Haiti: Breve história**. 2022. Disponível em: <https://bit.ly/3AzbO2Y>. Acesso em: 19 jan. 2022.

_____. **Twitter: Apresentação e uso**. 2022. Disponível em: <https://blog.twitter.com/en_us/topics/company>. Acesso em: 19 jan. 2022.

WIKIPEDIA. **Haiti: Apresentação**. 2022. Disponível em: <https://pt.wikipedia.org/wiki/Haiti>. Acesso em: 3 jan. 2022.

WIKIPEDIA. **Terremoto: Terremoto de 12 de Janeiro de 2010**. 2022. Disponível em: <https://fr.wikipedia.org/wiki/S%C3%A9isme_de_2010_en_Ha%C3%AFti>. Acesso em: 19 jan. 2022.