UNIVERSIDADE DE TAUBATÉ

Vanessa Morais Silver

PROTÓTIPO DE SISTEMA DE BUSINESS INTELLIGENCE VOLTADO PARA ANÁLISE DO IMPACTO DE NOTÍCIAS NA REDE SOCIAL

Taubaté
2019

UNIVERSIDADE DE TAUBATÉ

Vanessa Morais Silver

PROTÓTIPO DE SISTEMA DE BUSINESS INTELLIGENCE VOLTADO PARA ANÁLISE DO IMPACTO DE NOTÍCIAS NA REDE SOCIAL

Trabalho de Pós-Graduação Interdisciplinar apresentado como requisito parcial para a conclusão do curso de Gestão de Projeto de Business Intelligence do Departamento de Informática da Universidade de Taubaté.
Orientador: Prof. Dr. Luis Fernando de Almeida.

Taubaté
2019

Ficha catalográfica elaborada pelo
SIBi – Sistema Integrado de Bibliotecas / UNITAU

S587p Silver, Vanessa Morais
Protótipo de sistema de *Business Intelligence* voltado para análise do impacto de notícias na rede social / Vanessa Morais Silver. - 2019.
54f. :il.

Monografia (especialização) - Universidade de Taubaté, Departamento de Informática, 2019.
Orientação: Prof. Dr. Luís Fernando de Almeida, Departamento de Informática.

1. *Business Intelligence*. 2. Inteligência artificial. 3. Linguagem natural. 4. Redes sociais. 5. Análise de sentimentos. I. Universidade de Taubaté. II. Título.

CDD 006.3

Ficha catalográfica elaborada pela Bibliotecária Regina Márcia Cuba – CRB 8ª/7416

VANESSA MORAIS SILVER

PROTÓTIPO DE SISTEMA DE BUSINESS INTELLIGENCE VOLTADO PARA ANÁLISE DO IMPACTO DE NOTÍCIAS NA REDE SOCIAL

Trabalho de Pós-Graduação Interdisciplinar apresentado como requisito parcial para a conclusão do curso de Gestão de Projeto de Business Intelligence do Departamento de Informática da Universidade de Taubaté.

Data: _____

Resultado: _____

BANCA EXAMINADORA

Prof. Dr. Luis Fernando de Almeida Universidade de Taubaté

Assinatura_____

Prof. Dra. Rita de Cássia Rigotti Vilela Monteiro Universidade de Taubaté

Assinatura_____

Prof. Dr. José Carlos Lombardi Universidade de Taubaté

Assinatura_____

Dedico este trabalho de pós-graduação ao meu namorado Lucas, que sempre acreditou em mim, sendo essencial na minha vida e na realização deste sonho.

Vanessa Morais Silver

AGRADECIMENTOS

Agradeço primeiramente a Deus, a minha família que sempre me amparou e ficaram do meu lado, em especial aos meus pais, namorado, os meus amigos e professores que me ajudaram a conquistar esta grande etapa na minha vida.

Vanessa Morais Silver

RESUMO

PROTÓTIPO DE SISTEMA DE BUSINESS INTELLIGENCE VOLTADO PARA ANÁLISE DO IMPACTO DE NOTÍCIAS NA REDE SOCIAL

Com a difusão crescente das redes sociais associado ao crescimento de usuários conectados, surge um novo comportamento de divulgações de notícias, no qual os jornais tendem a serem digitais para fornecer na Web informações em tempo real. Entretanto, com o grande volume de informações e aumento das concorrências de fontes de jornais, ocasionam-se dúvidas de sobre quais fontes apresentam melhores resultados, qual o nicho de cada uma delas e qual é o perfil dos seus leitores. O protótipo proposto neste trabalho visa analisar o impacto de determinadas notícias na rede social Twitter, por meio de: conceitos de Business Intelligence; técnicas de Inteligência Artificial, como Processamento de Linguagem Natural e Análise de Sentimento; métodos de Web Scraping para captura de dados das notícias nas páginas Web; e a visualização dos dados por meio de indicadores. Com isso, têm-se como resultado métricas que mostram o comportamento dos usuários de determinadas notícias comparando com outras fontes, métricas de sentimentos nas publicações, o impacto gerado, temas mais comentados e preferências do público. Assim, é possível fornecer informações sobre comportamentos das notícias e de fontes de informações de maneira personalizada, inteligente e diária para o monitoramento dos resultados no setor jornalístico, auxiliando tomadas de decisões futuras.

Palavras-chave: Análise de Sentimento. Business Intelligence. Inteligência Artificial. Processamento de Linguagem Natural. Rede Social.

ABSTRACT

BUSINESS INTELLIGENCE PROTOTYPE PROTOTYPE FOCUSED ON SOCIAL NEWS IMPACT ANALYSIS

With the growing diffusion of social networks associated with the growth of connected users, a new behavior of news dissemination emerges, in which newspapers tend to be digital to provide real-time information on the Web. However, with the large volume of information and increased competition from newspaper sources, there are doubts about which sources have better results, what is the niche of each one and what is the profile of their readers. The prototype proposed in this paper aims to analyze the impact of certain news on the social network Twitter, through: Business Intelligence concepts; Artificial Intelligence techniques, such as Natural Language Processing and Sentiment Analysis; Web Scraping methods for capturing news data on Web pages; and data visualization through indicators. This results in metrics that show the behavior of certain news users compared to other online newspaper sources, sentiment metrics in publications, the impact generated, more talked about topics, and audience preferences. Thus, it is possible to provide information about news behaviors and information sources in a personalized, intelligent and daily way to monitor results in the news sector, helping to make future decisions.

Keywords: Sentiment Analysis. Business Intelligence. Artificial Intelligence. Natural Language Processing. Social Networking.

LISTA DE ILUSTRAÇÕES

LISTA DE QUADROS

LISTA DE ABREVIATURAS E SIGLAS

BI	*Business Intelligence*
IA	Inteligência Artificial
DW	*Data Warehouse*
PLN	Processamento de Linguagem Natural
OLTP	*On-line Transaction Processing*
OLAP	*On-line Analytical Processing*
KDD	*Knowledge Discovery in Databases*
VADER	*Valence Aware Dictionary for Sentiment Reasoning*

SUMÁRIO

1 INTRODUÇÃO

A difusão cada vez maior das redes sociais modifica-se a vida das pessoas e torna-se essenciais para grande parte da população. Junto com este crescimento de usuários conectados, há um grande volume de informações que são geradas e bombardeadas nas redes sociais a cada segundo, ocasionando um impacto de grande relevância nos usuários, difundindo conceitos, influenciando, estimulando e até modificando as escolhas do público.

Um exemplo das vantagens do uso das redes sociais pode ser observado nas divulgações de notícias, onde os jornais estão digitais e utiliza-se a *web* para fornecer informações em tempo real, e com isso, os usuários estão comentando sobre essas matérias nas mídias sociais, auxiliando nas divulgações e interagindo com as notícias em rede mundial.

Entretanto, o grande volume de informações e diversidades de públicos, gostos e afins, dificulta-se a criação de notícias que realmente impactam e alcançam as repercussões esperadas. Fora que a concorrência de fontes de jornais está aumentando, ocasionando dúvidas de quais apresentam melhores resultados e qual o nicho de cada uma delas.

Uma solução para o problema da falta de entendimento do impacto das notícias é a junção de *Business Intelligence* (BI) com técnicas de Inteligência Artificial (IA), como Processamento de Linguagem Natural (PLN) e Análise de Sentimentos, para mapear a relevância das notícias e monitorar as repercussões na rede social, identificando perfis, analisando as opiniões e inspecionando influências e resultados.

Portanto, a junção de BI com técnicas de IA visa minimizar erros de criações de notícias ruins e obsoletas, monitorar o impacto dessas notícias em um ambiente dinâmico, como a rede social, e transformar as fontes de notícias em um ambiente inteligente e capaz de se adaptar e conhecer seu leitor de forma independente, automática e simples.

1.1 DESCRIÇÃO DO PROBLEMA

A grande diversidade de fontes de informações dificulta o mapeamento de influência e importância do jornal virtual, ou seja, prejudica-se o levantamento dos resultados finais, porque não considera nas análises os meios de comunicações onde os leitores se manifestam e/ou pela fraca capacidade de diferenciar o real impacto dos resultados com de outras fontes. Além disso, a falta de conhecimento do perfil dos leitores, como eles estão reagindo e interagindo com as notícias e quais assuntos apresentam maior impacto, faz com que elas sejam criadas deficientemente ou com foco errado, podendo deixá-las obsoletas, não alcançando as repercussões esperadas.

1.2 OBJETIVO

1.2.1 Objetivo Geral

O objetivo desse trabalho de conclusão de curso é de criar um protótipo de *Business Intelligence* para analisar o impacto de determinadas notícias na rede social Twitter, a partir do monitoramento de relevância de fontes de informações, repercussões de notícias e tópicos, reações decorrentes e público alvo atingido.

1.2.2 Objetivos Específicos

Os objetivos específicos deste trabalho são:

- Estudar processo de *web scraping* em páginas *web* e a API da rede social Twitter;
- Armazenar dados históricos de notícias publicadas nas páginas selecionadas e de *tweets* publicados na rede social Twitter;
- Adotar conceitos de Business Intelligence no processo;

- Aplicar técnicas de Inteligência Artificial, como Processamento de Linguagem Natural e Análise de Sentimento;
- Desenvolver *dashboards* com *insigths* dos dados na ferramenta de visualização Microsoft Power BI.

1.3 RELEVÂNCIA

O trabalho proposto traz dados sobre as relevâncias de notícias e fontes de informações de maneira personalizada, inteligente e diária para o monitoramento dos resultados no setor jornalístico.

Além disso, mostra a Análise de Sentimentos das publicações das pessoas sobre as notícias, o impacto gerado, temas mais comentados e preferências do público, a partir do conceito de *Business Intelligence*, integrando dados capturados da rede social com fontes de notícias.

O trabalho vem como uma alternativa para análise de dados nas redes sociais, especificamente Twitter, para incentivar e divulgar o uso de técnicas de *Business Intelligence* com Processamento de Linguagem Natural e Análise de Sentimento, e assim tornar-se uma possibilidade robusta de manipulação dos dados gerados por meio de mídias sociais para levantamento de comportamento com as notícias publicadas.

1.4 ESTRUTURA DO TRABALHO

Este trabalho está dividido em cinco capítulos, sendo esse primeiro dedicado à introdução e visão geral do trabalho de pós-graduação.

Nos segundo e terceiro capítulos serão abordados a revisão bibliográfica e a metodologia, respectivamente.

Já no quarto capítulo será mostrado à solução proposta neste trabalho, processos e desenvolvimento do protótipo. No quinto capítulo será tratada a conclusão do trabalho de pós-graduação e as referências bibliográficas.

2 REVISÃO BIBLIOGRÁFICA

A revisão bibliográfica deste trabalho aborda os conceitos de *Business Intelligence*, *Data Warehouse*, Processamento de Linguagem Natural e Análise de Sentimento. Em seguida, trata-se a rede social Twitter, linguagem de programação python, técnicas de *web scraping* e as ferramentas Microsoft SQL Server e Microsoft Power BI.

2.1 BUSINESS INTELLIGENCE

As organizações armazenam um número cada vez maior de dados, pode chegar a ultrapassar *Petabytes* de dados retidos, e a velocidade dos negócios, aumento de concorrência, mudanças de cenários internos e externos das organizações forçam as respostas de dados serem em tempo real, precisa e informativa.

Sistemas transacionais OLTP (*On-line Transaction Processing*) não conseguem atender essas demandas, pois sua estrutura é definida para manipular operações cotidianas das empresas, utiliza-se bancos de dados relacionais e comandos de inserção, alteração, busca e exclusão dos dados. O OLTP supre as necessidades operacionais da empresa, e não auxilia as necessidades gerenciais em tempo real ou *online*.

Sistemas OLAP (*On-line Analytical Processing*) visa tratar grandes volumes de dados para análises de negócios, *insights* e suporte a tomadas de decisões. São sistemas que possuem um repositório com dados históricos, não voláteis e orientados a assuntos, com visualizações das informações gerenciais. Estes sistemas utilizam o conceito de *Business Intelligence* (BI) para criar todo o processo.

> "Os sistemas de *Business Intelligence* utilizam os dados disponíveis nas organizações para disponibilizar informação relevante para a tomada de decisão. Combinam um conjunto de ferramentas de interrogação e exploração dos dados com ferramentas que permitem a geração de relatórios, para produzir informação que será posteriormente utilizada pela

> gestão de topo das organizações, no suporte à tomada de decisão.” (Santos, 2006, p. 2)

De acordo com Figueira (1998), conceito de *Business Intelligence* utiliza o processo *Knowledge Discovery in Databases* (KDD) em suas etapas. Essas etapas são:

1. Entendimento da organização e suas estruturas: a primeira etapa do processo é entender como a organização funciona e como está estruturado para conseguir realizar os próximos passos;
2. Seleção dos dados: consiste em selecionar quais conjuntos de dados serão relevantes para construção do DW e extraí-los;
3. Pré-processamento e limpeza: etapa que consiste em realizar um pré-processamento dos dados para limpá-los, retirar todos ruídos, dados errôneos, ausentes, redundantes ou irrelevantes;
4. Transformação: responsável por tratar os dados para transformar em informações úteis e satisfatórias;
5. Mineração de dados: consiste em aplicar algoritmos de mineração para encontrar associações, padrões ou previsões e gerar informações.
6. Interpretação: responsável por interpretar e analisar os dados por meio de visualizações, relatórios, *dashboards* ou indicadores.
7. Conhecimento: etapa final do processo que resulta gerar conhecimento a partir de todas as etapas anteriores.

Ao empregar todo processo KDD, atinge-se o objetivo de gerar conhecimentos uteis para necessidades de nível gerencial, suprir demandas de análises de negócios, acompanhar e modelar dados históricos, metrificar dados e abranger todo conceito de *Business Intelligence*.

2.2 DATA WAREHOUSE

Data Warehouse (DW) é um armazém de dados, que contém todas as informações relevantes das organizações. Ele suporta grandes volumes de dados estruturados e não estruturados, apresenta uma arquitetura voltada para agilizar

consultas pesadas em tempo real ou *online*, armazena dados históricos e utiliza dimensionamento em suas tabelas.

O DW é responsável por solucionar os problemas que banco de dados de sistemas transacionais possui em não armazenar dados históricos, usar comandos de exclusão e alteração dos dados, arquitetura relacional entre tabelas, onde dificulta-se as otimizações das consultas e tempo de processamento para alcançar os resultados esperados de *Business Intelligence*.

Em uma organização, além do *Data Warehouse* que possui toda a informação da empresa, pode-se ter Data Marts para subconjuntos de dados do DWs referentes a determinado assunto afim de atender cada área individualmente.

A partir disso, de acordo com Machado (2000), um DW apresenta as seguintes características:

- Processo de criação: existem três tipos de processo de criação de um DW, sendo eles: *Top-Down*, *Bottom-Up* e a mistura dos dois anteriores. O primeiro tipo de processo de criação segue o modelo tradicional de *Data Warehouse* e depois a criação do *Data Mart*, já o segundo cria-se o *Data Mart* e em seguida o *Data Warehouse*.
- Construção por assuntos: O *Data Warehouse* é construído por assuntos, ou seja, ele é modelado ao redor dos assuntos principais da empresa para atender de forma mais efetiva as perguntas e solicitações de negócio. Utiliza-se para isso dimensões e fatos;
- Criação de dimensões: são tabelas que possuem dados descritivos para qualificar as informações das tabelas fatos;
- Criação de fatos: são tabelas que ligam várias tabelas dimensões e apresentam métricas, ou seja, dados quantitativos que cruzados com as dimensões e geram informações.
- Estrutura dimensional: a estrutura dimensional padrão utilizada em projetos de DW são o esquema estrela, na qual as dimensões são ligadas somente nos fatos. Pode-se uma dimensão possuir ligações em mais de uma fato, porém fatos não apresentam ligações entre si.
- Dados não voláteis e históricos;

Com essas etapas e características, o *Data Warehouse* se torna fundamental para Business Intelligente, sendo uma das principais bases do processo.

2.3 PROCESSAMENTO DE LINGUAGEM NATURAL

Segundo Loper, Bird e Klein (2009), linguagem natural são as comunicações cotidianas entre os humanos, diferenciando, por exemplo, das linguagens artificiais utilizadas para se comunicar e interagir com os computadores, como linguagens de programação. O processo de manipular essas linguagens naturais é definido como Processamento de Linguagem Natural.

Processamento de Linguagem Natural (PLN) é responsável por fazer o computador compreender uma sentença, interpretar e gerar informações úteis da linguagem humana, a partir de métodos de inteligência artificial.

Os métodos de PLNs apresentam várias técnicas de manipulação textual, na qual combinados satisfazem diversos objetivos na área de linguagem natural. De acordo com Loper, Bird e Klein (2009), esses métodos são:

- Normalização: responsável por limpar a sentença, retirar caracteres especiais e/ou caracteres alfanuméricos, espaços em branco em excesso, *tags* de marcação e conversão de todas as palavras para letra maiúscula ou minúscula.
- Tokenização lexical: os métodos de tratamento de PLN não conseguem assimilar textos inteiros de uma única vez, para solucionar este problema deve-se quebrar as sentenças por frases ou por palavras (*tokens*) para a análise ser individual, dependendo do nível de detalhe a ser alcançado.
- *Stopwords*: responsável pela remoção de *stopwords*, que são as palavras irrelevantes do texto, como artigos e preposições. Exemplo: a, o, de, com, para.
- Stemização: os métodos de stemização são responsáveis por diminuir cada palavra para seu radical, assim, por exemplo, é possível manipular as palavras sem interferência de desinências verbais e nominais, ou seja,

é transformar a palavra em sua forma básica, sendo um radical um morfema indivisível e comum a um conjunto de palavras. Exemplo: Gato, gatas, gatinho são diminuídos em “gat”.

- Lematização: os métodos de lematização transformam as palavras em sua forma singular e masculina. Com isso, tem-se a redução do vocabulário e a abstração das palavras. Exemplo: Gato, gatas, gatinho são transformados em gato.
- Frequência: responsável por trazer o número de ocorrências de cada palavra em um documento. Neste método há a frequência inversa do documento para encontrar quais palavras que o caracteriza, frequência relativa para encontrar o número de ocorrências pelo tamanho do documento e frequência absoluta que conta as presenças das palavras.

Cada método apresentado pode-se ser combinado para conseguir um resultado satisfatório, sendo a normalização e tokenização indispensáveis para devido funcionamento dos outros métodos.

As técnicas empregadas nos métodos são diversas, pode-se usar algoritmos de rede neurais para treinar um modelo a identificar as palavras e tratá-las, um dicionário linguístico de base para consultas, tratamentos e análises contextuais, e junções de funções computacionais em gerais para criar algoritmos robustos.

O campo de PLN expande e melhora constantemente, emprega-se novos métodos e técnicas, aperfeiçoa-se processos e componentes, visando diminuir o espaço de divergência entre linguagens naturais e computadores.

2.4 ANÁLISE DE SENTIMENTO

Análise de sentimento é uma técnica de IA capaz de descobrir informações abstratas de uma frase ou texto; está sendo cada vez mais utilizada dentro de análises de dados das redes sociais, pois decifrar o que as pessoas estão escrevendo nas mídias sociais e medir a intensidade dos textos são fundamentais para mapeamento de dados.

> “O principal objetivo da análise de sentimentos é definir técnicas automáticas capazes de extrair informações subjetivas de textos em linguagem natural, como opiniões e sentimentos, a fim de criar conhecimento estruturado que possa ser utilizado por um sistema de apoio ou tomador de decisão.” (Benevenuto, 2015, p. 2)

Dessa maneira, a análise de sentimento utiliza a área de Processamento de Linguagem Natural para conseguir atingir seu objetivo. Emprega-se os métodos de PLN para normalizar, tokenizar, stemizar ou lematizar os textos, com o objetivo de preparar, tratar e classificar as palavras das sentenças para abstrair, facilitar o entendimento e aumentar a acurácia da análise de sentimento.

Com isso, é possível inserir um algoritmo de análise de sentimento no texto para medir a polaridade da sentença, podendo ser categórica (positiva, negativa, neutra) ou numérica. A polaridade numérica é o resultado de uma análise que corresponde o grau de sentimento de uma frase, no qual o valor retornado varia na faixa de 1 à -1, onde 1 significa positivo, 0 significa neutro e -1 significa negativo.

Para medir o valor da polaridade leva-se em conta vários fatores em uma sentença, como:

- Palavras empregadas e seus significados. Exemplo: bom e mal;
- Intensidade das palavras. Exemplo: “péssimo” apresenta uma intensidade maior que “ruim”;
- Gírias e emojis, muito encontrado em publicações nas redes sociais;
- Ênfases com letras maiúsculas. Exemplo: “ÓTIMO” apresenta maior ênfase e intensidade que “ótimo”;
- Pontuações, onde dependendo da pontuação e a quantidade repetida equivale a um sentimento.

Com os resultados da polaridade é possível compreender os textos, encontrar padrões, transformar sentenças em resultados quantitativos ou classificatórios para análises de comportamento e até prever futuros textos ou ações decorrentes das análises, sendo a rede social o principal ambiente para se empregar algoritmos de análise de sentimento treinados pelo número significativo de textos publicados a cada segundo.

2.5 SEMELHANÇA DE COSSENO

Técnicas de cálculos de semelhanças de documentos utilizam métodos de PLN tanto para preparar o conteúdo como para calcular a métrica. Para preparar os textos, realiza-se limpeza dos dados, padronização de escrita, tokenização por palavras, remoção de stopwords, lematização ou stematização. Com isso, evita-se desvios nas análises de semelhanças provocadas pelas concordâncias gramaticais.

Semelhança de Cosseno é uma das métricas existentes para cálculo de similaridade entre dois textos, que, segundo Sidorov (2014), a métrica utiliza o modelo de espaço vetorial, o qual faz uso do cosseno como medida tradicional, para determinar a semelhança entre dois objetos representados como vetores, ou seja, o modelo mede o cosseno do ângulo entre dois vetores inseridos no espaço multidimensional.

Assim, a Semelhança de Cosseno consegue identificar o grau de similaridade entre documentos pelo assunto abordado, sendo o tamanho dos documentos irrelevantes, pois quanto maior um documento mais chances ele tem de possuir as mesmas palavras que outros, mesmo sendo de assuntos divergentes.

O processo de cálculo da Semelhança de Cosseno considera cada palavra de um documento como dimensão em um espaço multidimensional e trabalha com o ângulo dos documentos em vez da magnitude da distância euclidiana que considera o tamanho dos documentos, ou seja, quanto maior a semelhança dos assuntos, menor o ângulo entre eles.

Dessa maneira, segundo Sidorov (2014), para conseguir a métrica da Semelhança de Cosseno, deve-se:

1. Contar as frequências de palavras de cada documento a ser analisado;
2. Calcular o resultado pela equação abaixo, considerando dois vetores de texto a e b:

$$cosine(a,b) = \frac{a \cdot b}{||a|| \times ||b||} \quad (1)$$

$$cosine(a,b) = \frac{\sum_{i=1}^{N} a_i b_i}{\sqrt{\sum_{i=1}^{N} a_i^2}\sqrt{\sum_{i=1}^{N} b_i^2}}. \quad (2)$$

A métrica de Semelhança de Cosseno pode ser aperfeiçoada inserindo métodos de Inteligência Artificial para criar modelos treinados que conseguem entender quais palavras são semelhantes, independente da escrita. Exemplo: Olá e oi.

2.6 REDE SOCIAL TWITTER

Rede social é uma plataforma web voltada para relações sociais de pessoas, organizações e eventos, conectados em tempo indeterminado e ilimitado com compartilhamentos de vários tipos de informações, interações e comunicações.

A rede social Twitter é uma plataforma de informações instantâneas criada em 2006 para interações de publicações de vários formatos, com vários tipos de conexões e funcionalidades.

Atualmente, ela possui milhões de usuários cadastrados no mundo inteiro que publicam e comentam a todo momento por meio de *tweets*, que são as publicações realizadas por eles.

Além disso, a rede social apresenta uma API própria para desenvolvedores, onde disponibiliza comandos para extrações de dados na sua plataforma por meio de *token*s de acesso. Os *tokens* devem ser solicitados pela própria plataforma e utilizados assim que liberado os acessos em qualquer *software* ou programa.

2.7 PYTHON

Algoritmos são instruções criadas para o computador seguir e realizar determinados objetivos.

> “Informalmente, um algoritmo é qualquer procedimento computacional bem definido que toma algum valor ou conjunto de valores como entrada e produz algum valor ou conjunto de valores como saída. Portanto, um algoritmo é uma sequência de passos computacionais que transformam a entrada na saída.” (Cormen, 2002, p. 21)

Os algoritmos são construídos através de linguagens de programação, que são linguagens escritas com regras de sintaxes, processamentos e execuções próprias.

Atualmente, existem várias linguagens de programação com propósitos diversos, sendo classificadas por linguagem de baixo e alto nível.

Linguagens de baixo nível são instruções voltadas diretamente para a arquitetura do computador, sendo de difícil compreensão humana. Já as linguagens de alto nível são instruções abstratas e de fácil entendimento, na qual precisam de um compilador para transformar as instruções compreensíveis para o computador.

A linguagem de programação Python é considerada de alto nível, voltada para desenvolvimento de algoritmos de ciência de dados e aplicações web, criada em 1991. Ela é uma linguagem de código aberto e grátis, contendo um elevado número de *frameworks* disponíveis.

2.8 WEB SCRAPING

Segundo Mitchell (2015), *web scraping* é a prática de coleta dos dados por meio de um programa computacional integrado a uma API, sendo utilizado, principalmente, em consultas na *web*.

O processo de *web scraping* engloba técnicas de programação para a extração, tratamento e carga dos dados de maneira automática, lendo e capturando informações HTML e arquivos contidos nas páginas *web*.

A partir desse processo, a captura das informações é muito mais rápida, acessível e abrangente, podendo chegar a milhares de dados coletados por dia.

2.9 MICROSOFT SQL SERVER E MICROSOFT POWER BI

Microsoft SQL Server é um sistema gerenciador de banco de dados desenvolvido pela Microsoft, com recursos de segurança, controle e logs de todos os processos, comandos SQL otimizados, ambiente de fácil manuseio, suporte a banco de dados transacionais, *Data Warehouse*, *Data Marts* e *Big Data*.

Já o Microsoft Power BI é um *software* voltado para *Business Intelligence*, contendo serviços de análises de negócios para gerar informações de insights e entendimento dos dados.

Ele fornece opções para extração, tratamento e carga dos dados, integração com outras aplicações da Microsoft, conexões com diversos Bancos de Dados e estrutura para *Data Warehouse*, além de um número considerável de opções para criações visuais de relatórios e *dashboards*.

3 METODOLOGIA

Em suma o desenvolvimento deste trabalho consistiu nas seguintes etapas:

- Definição do escopo e limitação do problema: A definição do escopo baseia-se no levantamento do problema inicial do protótipo, onde apresenta uma oportunidade de melhoria no processo de monitoramento e análises de notícias em ambiente virtual. Assim, o escopo visa contribuir para melhorar as análises e medições dos dados gerados pelas notícias no setor jornalístico e abordar técnicas de IA e BI para criar um monitoramento que integre o ambiente das fontes primárias das matérias com o ambiente final dos leitores nas mídias socias. A limitação do problema se baseia na falta de técnicas capazes de intermediar os jornais virtuais com redes sociais de maneira independente e personalizada.
- Levantamento bibliográfico sobre técnicas a serem utilizadas: Para o desenvolvimento do presente trabalho utilizam-se pesquisas bibliográficas que se baseiam em publicações científicas da área de BI e as técnicas de IA de PLN e Análise de Sentimento. Já o estudo de caso engloba o processo de tratamento de PLN e as metodologias de BI, seguindo o processo KDD na construção do protótipo.
- Análise e definição das fontes de dados de notícias a serem utilizadas: Com relação aos dados, foram utilizadas as fontes primárias de dados dos jornais virtuais G1 e Folha para extrair as cinco primeiras notícias das categorias: últimas notícias publicadas e manchetes. As fontes de jornais virtuais selecionadas apresentam alto número de visualizações, sendo uns dos principais jornais virtuais do Brasil. Além disso, para realização de *web scraping*, as páginas web desses jornais apresentam padrões e tags que facilitam e permitem a extração de informações.
- Análise e definição da rede social a ser utilizada: foram utilizados os dados dos *tweets* publicados na rede social Twitter para análises, a partir de filtros de busca pelas cinco principais palavras chaves e título

da notícia a ser analisada, sendo estas etapas processadas todos os dias no período da noite. Emprega-se a rede social Twitter por ela possuir: grandes volumes de informações e interações entre usuários e fontes de jornais virtuais; API própria de extração de dados para linguagem Python; solicitações de *tokens* simples e rápidas e possibilidade de criar processos automáticos de extração diária de dados. Já outras redes sociais, como Facebook e Instagram, foram descartadas por limitar acessos e possuir processos de solicitações de *tokens* extensos e burocráticos, com baixas taxas de permissão de acesso e de resposta rápida.

- Implementação do módulo plataforma web: O protótipo é uma plataforma web, na qual utiliza-se ferramentas de implementação Django, sendo um framework voltado para desenvolvimento de sites na linguagem de programação Python, devido a facilidade de implementação da linguagem, rápido desempenho e processamento dos dados, integrações com outras APIs de forma simplificada e diversos recursos para manipulação de dados voltados para Business Intelligence e ciência de dados. Foram utilizadas bibliotecas padrões da linguagem Python e o Bootstrap, para criar interfaces responsivas, personalizadas e simplificadas.
- Implementação do módulo Inteligência Artificial: foram utilizados os frameworks Python NLTK e SpaCy de Processamento de Linguagem Natural, para o tratamento e manipulação de notícias e *tweets*; o VADER (*Valence Aware Dictionary for Sentiment Reasoning*) para realizar a Análise de Sentimento, na qual utiliza regras léxicas e regras sintáticas no seu algoritmo voltado para análises de textos de mídias sociais.
- Implementação do módulo visualizações: emprega-se o Microsoft Power BI, software de visualizações e análises de BI, voltado para gerar conhecimentos e insights para ajudar tomadas de decisões e gestão por meio de *dashboards*. O software contém versão grátis para estudantes e opções de compartilhamentos de conteúdo e integrações com bancos de dados.

- Implementação do módulo *Data Warehouse*: o *Data Warehouse* é um banco de dados que usa a Microsoft SQL Server para a criação, domínio, manuseio de dados por meio de tabelas de dimensões e fatos. Apresenta-se compatibilidade com outros produtos da Microsoft, como o Microsoft Power BI, utilizado no protótipo.
- Testes do protótipo: os testes do protótipo são voltados em verificar os conteúdos extraídos dos jornais virtuais, a qualidade de informações que são armazenadas no *Data Warehouse* da rede social Twitter, a medição da confiabilidade dos resultados dos métodos de Semelhança de Cosseno e da Análise de Sentimento, a eficácia no tratamento da técnica de PLN, conferindo as interpretações e resultados obtidos dos textos das notícias, e os testes realizados nas manipulações e interações dos *dashboards*.
- Análise dos resultados: as análises dos resultados envolvem em identificar o desempenho dos jornais virtuais comparando os números de publicações na rede social, verificar relevâncias das repercussões dos segmentos das notícias, identificar padrões de comportamentos e análises de sentimentos em informações das publicações, identificar e metrificar os resultados da similaridade das notícias e o grau de exclusividade, analisando os impactos por regiões.

4 DESENVOLVIMENTO

O protótipo proposto neste trabalho desenvolve-se na linguagem de programação Python, com a utilização do framework Django para ambiente *web*. As etapas de desenvolvimento baseiam-se no processo KDD no conceito de *Business Intelligence*.

Com isso, este capítulo está divido em: mapeamento do ambiente; extração dos dados das notícias; pré-processamento, limpeza e transformação das notícias; carga das notícias; similaridade de texto; monitoramento do Twitter; extração dos dados no Twitter; pré-processamento, limpeza e transformação dos dados do Twitter; carga dos dados do Twitter; dicionário de dados; *Data Warehouse* e *dashboards*, respectivamente.

4.1 MAPEAMENTO DO AMBIENTE

O protótipo visa analisar as repercussões e métricas das notícias na rede social. Para isso, mapeia-se os ambientes que proporcionam estas informações para conseguir alcançar o objetivo traçado.

O primeiro ambiente corresponde aos dados das notícias que se encontram em jornais virtuais na *web*, por meio de páginas HTML, que possibilitam a extração de informações instantânea e automática. Com isso, seleciona-se os jornais virtuais G1 e Folha, na qual possuem um padrão de estrutura HTML e um considerável nível de visualizações.

As informações destes *sites* são publicadas no decorrer do dia e contém data de publicação, categorias, títulos e etc. Os *sites* possuem um padrão de ter uma página inicial com todos os anúncios das notícias e, ao selecionar uma, redireciona-se para a todo o conteúdo e detalhes dela.

Estes *sites* virtuais apresentam segmentos de classificações de acordo com o comportamento da notícia. Utiliza-se neste protótipo os segmentos de manchetes e últimas notícias publicadas para extrair as cinco primeiras de cada uma.

A partir dessas informações, realiza-se o processo de pré-processamento, limpeza e transformação para padronizar, limpar a preparar os dados para monitoramento de notícias e análises textuais. As análises textuais atuam nos textos de cada notícia para encontrar a frase mais relevante e as cinco principais palavras-chaves do texto; utiliza-se estas informações para buscas no Twitter, comparativo de semelhança das notícias e predominância de assuntos nos textos.

O segundo ambiente corresponde aos dados dos *tweets* do Twitter. Esses *tweets* são publicados pelos usuários a todo momento na rede social, na qual se manifestam de acordo com suas opiniões pessoais e gostos. É possível trabalhar com texto, localização, data de publicação e interações com outros usuários da rede em cada *tweet* extraído, além de possibilitar buscas por frases, textos ou palavras-chaves.

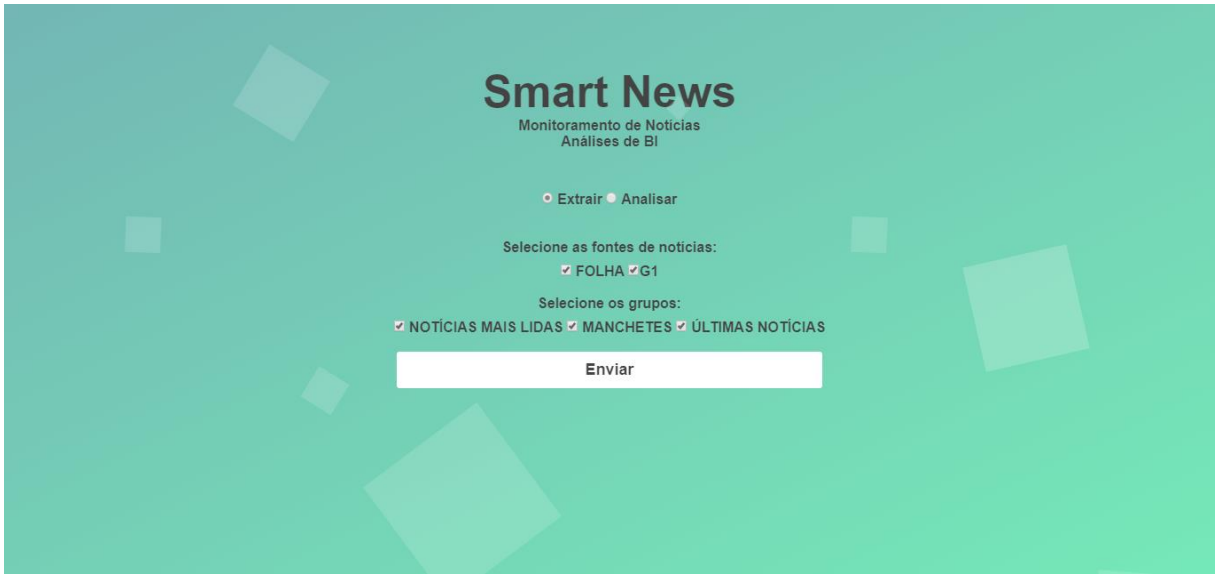
Assim, as informações estão disponíveis e atualizadas diariamente nos dois tipos de ambientes abordados no protótipo, sendo os dados semi-estruturados, ou seja, os dados não estão em um formato estruturado para ser armazenado diretamente em um banco de dados, porém apresentam *tags* e padrões que possibilitam a separação de cada informação a ser trabalhada.

4.2 EXTRAÇÃO DOS DADOS DAS NOTÍCIAS

Para a extração das informações das notícias, usa-se os recursos da linguagem de programação Python para realizar o *web scraping* dos dados

relevantes. Todas as etapas de extração processam-se de maneira automática, sem precisar da intervenção humana, basta o usuário selecionar qual jornal virtual e categorias (últimas notícias e manchetes) ele quer realizar a extração, por meio do portal *web* visto na Figura 1, ou esperar o processo agendado default ser executado todos os dias às 17hs.

Figura 1 – Tela de Extração



Fonte: Autora

Para a extração ser executada, armazena-se no *Data Warehouse* os *links* de acessos dos *sites* virtuais selecionados, assim o protótipo consegue identificar quais ele deve navegar para conseguir os dados, ou seja, mapeia-se todos os *links* que devem ser extraídos.

Com isso, realiza-se as seguintes etapas no processo de extração de dados das notícias:

1. Verificação de quais fontes de dados e quais categorias extrair, de acordo com os valores da seleção do usuário ou do processo default. O processo default extrai dados de todas as fontes e categorias no sistema;
2. Navegação nas páginas dos jornais virtuais de acordo com os *links* de acessos registrados no *Data Warehouse*;
3. Extração dos títulos e *links* dos anúncios das cinco primeiras notícias de cada categoria dos jornais;

4. Navegação na página de cada anúncio das notícias do passo anterior para extração das informações detalhadas: título, texto, data de publicação e segmento de cada notícia.

Processa-se as etapas de extração em alguns minutos, lendo e capturando as informações relevantes de cada página HTML.

4.3 PRÉ-PROCESSAMENTO, LIMPEZA E TRANSFORMAÇÃO DAS NOTÍCIAS

Após a extração, é necessário fazer um pré-processamento para limpar e transformar os dados. Retira-se os ruídos e sujeiras das informações, transforma-se em dados relevantes para carregar no *Data Warehouse*.

As etapas destes processos são:

1. Pré-processamento das notícias extraídas para as próximas etapas;
2. Limpeza dos títulos das notícias: responsável por remover espaços em branco em excesso e remover caracteres especiais ou palavras que se encontram antes ou depois do título;
3. Limpeza e transformação dos textos das notícias: por meio de técnicas de Processamento de Linguagem Natural realiza-se a remoção de pontuações, conversão de todo o texto para letra minúscula, tokenização por palavras, remoção de *stopwords* e lematização de todo conteúdo;
4. Transformação das datas de publicação da notícias: responsável por remover espaços em branco em excesso e padronizar as datas e horas em um único formato. Exemplo: 12/09/2019 08:00.

Em seguida, com os dados das notícias limpas e transformadas pelas técnicas de PLN, realiza-se o processos de análises textuais logo abaixo:

- Top cinco das palavras-chaves: responsável por encontrar as cincos palavras-chaves do texto das notícias por meio de técnicas de Processamento de Linguagem Natural, calcula-se a frequência absoluta e relativa das palavras, sendo elas a contagem de frequência de cada

palavra no texto e a divisão da frequência absoluta por total de palavras, respectivamente. Em seguida, mede-se a relevância de cada palavra ordenando-se em ordem decrescente as duas frequências. Assim, as cinco primeiras palavras-chaves são as que mais representam o texto analisado.

- Frase mais relevante: responsável por encontrar a frase mais relevante do texto das notícias por meio de técnicas de Processamento de Linguagem Natural para realizar a tokenização de frases e calcular um score delas por meio das frequências das palavras. A partir dos resultados, a frase com maior número de score é a frase mais relevante.

O processo de pré-processamento, limpeza e transformação visa modificar as informações extraídas em dados significativos para o objetivo do protótipo e gerar por meio de análises textuais dados complementares ao conjunto extraído.

4.4 CARGA DAS NOTÍCIAS

Neste processo, realiza-se a carga de notícias limpas e transformadas no *Data Warehouse* em uma tabela dimensão correspondente para manter o histórico e dimensionamento de todas as notícias capturadas no decorrer do tempo, na qual pode ser observada no item 4.12 dedicado as tabelas do banco de dados.

4.5 SIMILARIDADE DE TEXTO

Todas as notícias carregadas no *Data Warehouse* passam por uma mineração de texto para encontrar a sua similaridade.

Similaridade de texto visa metrificar o quanto uma notícia é semelhante a outras, tanto do mesmo jornal virtual como de outros. Assim, é possível analisar o

grau de relevância e exclusividade das notícias já publicadas, e verificar o quanto outros *sites* estão abordando o mesmo tema.

Para alcançar esta finalidade, os textos de cada notícia precisam estar lematizados, pois, para calcular a similaridade, as palavras no texto precisam estar no formato singular e masculino para evitar desvios causados pela concordância gramatical. Este método de lematização de Processamento de Linguagem Natural efetua-se na etapa de pré-processamento, limpeza e transformação dos dados das notícias.

Assim, para encontrar a similaridade do texto, executa-se os seguintes passos para cada notícia carregada no *Data Warehouse*:

1. Busca-se todas as notícias cadastradas no *Data Warehouse* no período de cinco dias atrás para análise. Usa-se o período de cinco dias, pois elas apresentam alto nível de rotatividade e, após este período, os dados se tornam irrelevantes para este tipo de análise.
2. Agrupa-se as notícias de acordo com o jornal virtual extraído e exclui-se as repetidas do mesmo jornal.
3. Cria-se matrizes vetorizadas dos textos para cálculo da média da Semelhança de Cosseno de cada grupo.
4. Calcula-se a relevância do texto da notícia por meio da subtração dos resultados encontrados do passo 3 de cada grupo, conforme equação (3).

$$\text{Relevância} = X - Y \qquad (3)$$

Onde,

X = média da Semelhança de Cosseno do jornal virtual da notícia analisada.

Y = média da Semelhança de Cosseno de outros jornais virtual da notícia analisada.

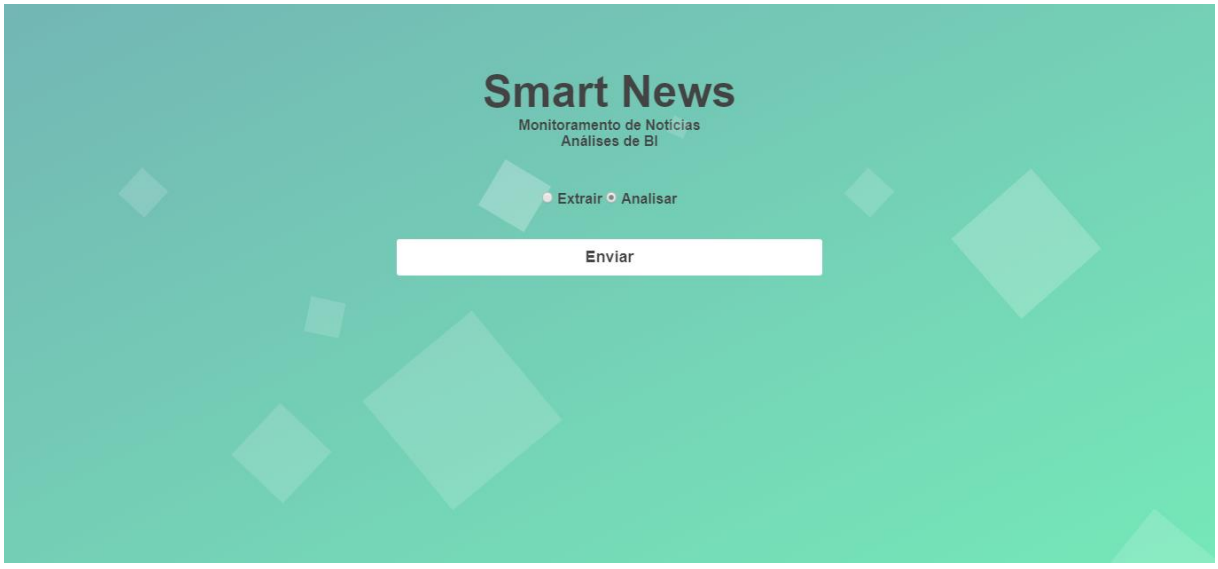
Após encontrar o resultado da similaridade de texto de cada notícia extraída, realiza-se o carregamento dessas informações na tabela fato correspondente no *Data Warehouse* (vide item 4.12).

4.6 MONITORAMENTO DO TWITTER

A partir das notícias armazenadas no *Data Warehouse*, realiza-se um monitoramento de seus dados na rede social Twitter diariamente para buscar publicações e interações dos usuários sobre elas em período de 1 semana, ou seja, notícias extraídas a partir de oito dias atrás não são mais analisadas, pois o tempo de vida de interações de usuários na rede social sobre elas são muito curtos, devido á alta rotatividade de novas notícias publicadas todos os dias.

Esse processo pode-se ser solicitado a qualquer momento no portal *web*, conforme Figura 2, ou aguardar o procedimento automático agendado para ser executado todos os dias as 17hs.

Figura 2 – Tela de monitoramento do Twitter



Fonte: a autora

Para o monitoramento, é necessário a utilização de *tokens* de acesso para conectar na API do Twitter. Os *tokens* são chaves de segurança disponibilizadas pela rede social ao requerimento de permissão de acesso na API. Assim, com ele, conecta-se e manipula-se todos os tipos de funcionalidades que ela possui, através de respostas de arquivos em JSON.

4.7 EXTRAÇÃO DOS DADOS NO TWITTER

A extração dos dados no Twitter inicia-se buscando todas as notícias armazenadas no *Data Warehouse* no período de uma semana para análise e, em seguida, conecta-se na API com o *token* de acesso.

Para cada notícia a ser analisada, busca-se os dados a ser extraídos por:

- Top 5 palavras-chaves: todas as notícias apresentam no *Data Warehouse* as cinco principais palavras-chaves que as caracterizam. Estas palavras são passadas para a API do Twitter para busca de *tweets,* por meio de buscas avançadas que trazem conteúdos que contenham todas as palavras-chaves pesquisadas.
- Título das notícias: a busca pelo título das notícias procura a frase exata nos *tweets*.
- Links das notícias: pelo grande número de usuários que utilizam somente os *links* das notícias nas publicações, realiza-se buscas pelos links nos *tweets*.

A partir das buscas, a API retorna os dados de todos os *tweets* encontrados em um arquivo JSON. Para extrair desse arquivo as informações relevantes, executam-se os passos abaixo:

1. Percorre-se todo arquivo e separa as informações de cada *tweet*;
2. Para cada informação extrai-se: id de publicação do *tweet*, texto, data de criação, quantidade de pessoas que gostaram e compartilharam o *tweet,* hashtags utilizadas e a localização que o *tweet* foi publicado.

4.8 PRÉ-PROCESSAMENTO, LIMPEZA E TRANSFORMAÇÃO DOS DADOS DO TWITTER

Ao finalizar a etapa de extração, executa-se o pré-processamento dos dados e as etapas de limpeza e transformação para armazenamento no *Data Warehouse*. As etapas são:

- Transformação das datas de criação dos *tweets*: responsável por remover espaços em branco em excesso e padronizar as datas e horas em um único formato. Exemplo: 08/06/2019 13:40.
- Limpeza de *hashtags*: método para remover espaços em branco em excesso, retirar caracteres que não fazem parte do *hashtags* e padronizar em único formato. Exemplo: #Economia.
- Transformação da localização: responsável por transformar em latitude e longitude a localização do *tweet*.

Com as etapas anteriores, os dados apresentam-se limpos e tratados, convertidos de informações semiestruturadas para dados padronizados, sem ruídos e precisos.

4.9 ANÁLISE DE SENTIMENTO DOS DADOS DO TWITTER

Com os dados estruturados, realiza-se a análise para metrificar os sentimentos das publicações em todos os *tweets* extraídos.

Para isso, utilizam-se recursos da linguagem Python e o *framework* VADER para realizar a Análise de Sentimento, usando regras léxicas e regras sintáticas voltadas para mídias sociais, sendo necessário somente traduzir o texto a ser analisado para Inglês, linguagem nativa do *framework*.

O processo de análise do *framework* utiliza os recursos padrões de análise de sentimento para encontrar a pontuação positiva, negativa e neutra, ou seja, o quanto um texto tem de cada sentimento. Assim, utilizando a pontuação dos três sentimentos, calcula-se a pontuação geral da análise de sentimento do texto.

4.10 CARGA DOS DADOS DO TWITTER

Realiza-se a etapa de carga dos dados extraídos, processados e analisados do Twitter no *Data Warehouse*, em uma tabela fato correspondente para manter o histórico, dimensionamento e métricas de *tweets* capturados no decorrer do tempo, na qual pode ser observado no item 4.12 dedicado as tabelas do banco de dados.

4.11 DICIONÁRIO DE DADOS

O dicionário de dados do protótipo proposto neste trabalho apresenta as definições e estruturas de todas as informações contidas no *Data Warehouse*, sendo elas como dicionário na Quadro 1.

Quadro 1 – Dicionário de Dados

Campo	Tipo	Descrição
Id_newspaper	Chave primária inteira	Chave primária das fontes de notícias.
Name_newspaper	String	Nome das fontes de notícias. Ex.: G1, Folha.
Link_newspaper	String	Link de acesso às páginas das fontes de notícias.
Id_group	Chave Primária inteira	Chave primária dos agrupamentos das categorias das fontes de notícias.
Name_group	String	Nome dos agrupamentos das categorias das fontes de notícias. Ex.: Manchete.
Id_ news_publication	Chave primária inteira	Chave primária das notícias carregadas.
Data_ref_ news_publication	Data	Data de referência de carga da notícia no *Data Warehouse*.
Title_ news_publication	String	Título da notícia.
Text_ news_publication	String	Texto da notícia.
Link_ news_publication	String	Link de acesso da página da notícia.
Key_words_ news_publication	String	Top 5 palavras-chaves da notícia.
Key_phase_ news_publication	String	Frase mais significativa da notícia.
Id_segment	Chave primária inteira	Chave primária dos segmentos das notícias carregadas.
Name_segment	String	Nome do segmento da notícia. Ex.: Política.

Id_ news_relevance	Chave primária inteira	Chave primária das métricas de relevância das notícias de Semelhança de Cosseno.
Data_ref_ news_relevance	Date	Data de referência da carga da relevância da notícia no Data Warehouse.
simulation_score_x_ news_relevance	Float	Valor da Semelhança de Cosseno da notícia com o mesmo jornal virtual.
simulation_score_y_ news_relevance	Float	Valor da Semelhança de Cosseno da notícia com outros jornais virtuais
relevance_score news_relevance	Float	Valor final da Semelhança de Cosseno da notícia.
Id_tweet	Chave primária inteira	Chave primária dos dados do *tweet*.
Data_ref_tweet	Date	Data de referência da carga do *tweet* no *Data Warehouse*.
Option_tweet	Integer	Opção de tipo de extração do *tweet*: palavras-chaves, link ou títulos.
Cod_tweet	Integer	Código do *tweet* no Twitter.
Text_tweet	String	Texto da publicação do *tweet*.
Data_publication_tweet	Date	Data de publicação do *tweet* no Twitter.
Count_likes	Integer	Quantidade de curtidas no *tweet*.
Count_retweets	Integer	Quantidade de compartilhamentos no *tweet*.
Hashtags	String	Hashtags publicados no *tweet*.
Location	String	Localização geográfica da publicação do *tweet*.
Latitude	Float	Valor da latitude da localização geográfica da publicação do *tweet*.
Longitude	Float	Valor da longitude da localização geográfica da publicação do *tweet*.
Count_tweet	Integer	Quantidades de caracteres no texto do *tweet*.
Negative_score	Float	Valor negativo da análise de sentimento do texto do *tweet*.
Positive_score	Float	Valor positivo da análise de sentimento do texto do *tweet*.
Neutral_score	Float	Valor neutro da análise de sentimento do texto do *tweet*.
Overall_score	Float	Valor geral da análise de sentimento do texto do *tweet*.

4.12 DATA WAREHOUSE

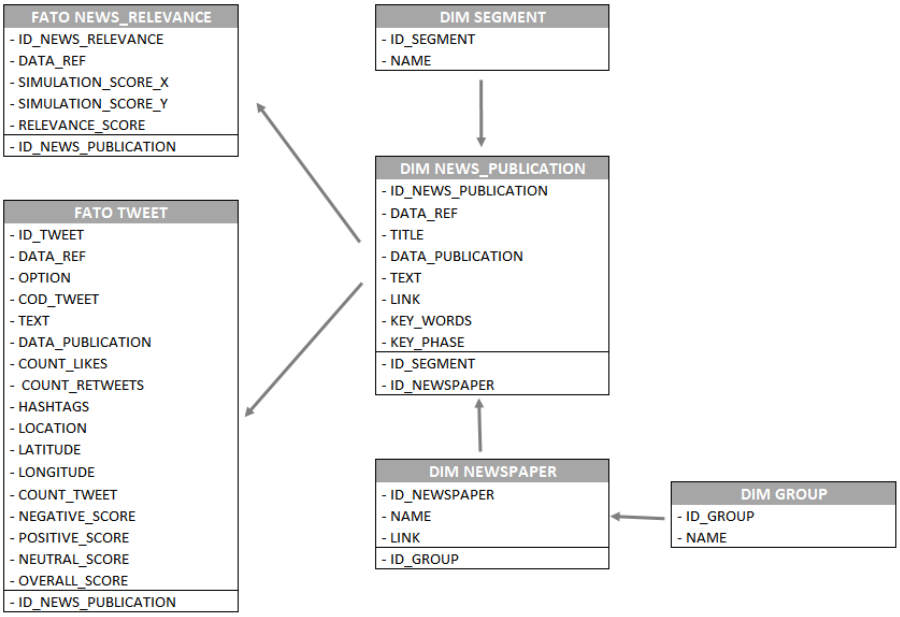
O *Data Warehouse* proposto neste trabalho, apresenta todas as características padrões de um projeto de *Business Intelligence* baseado nas boas práticas da área. Possui-se dimensionamento de tabelas, com fatos e dimensões, histórico de dados e opções somente de leitura. As tabelas são:

- Dimensão Segment: Tabela responsável por armazenar todos os nomes dos segmentos das notícias extraídas. Exemplo: Economia.

- Dimensão Group: Tabela responsável por armazenar todos os nomes dos grupos das categorias das notícias extraídas. Exemplo: Manchete, Últimas notícias.
- Dimensão Newspaper: Tabela responsável por armazenar o nome e link de todos os jornais virtuais.
- Dimensão News_publication: Dimensão responsável por armazenar todas as informações relevantes das notícias extraídas.
- Fato News_Relevance: Tabela fato responsável por armazenar as métricas referentes as relevâncias de todas as notícias armazenadas no *Data Warehouse*.
- Fato *Tweet*: Tabela fato responsável por armazenar as informações e métricas referentes aos dados de todos os *tweets* extraídos.

Assim, com as tabelas dimensões e fatos, o *Data Warehouse* possui o diagrama UML da Figura 3.

Figura 3 – Diagrama UML



Fonte: a autora

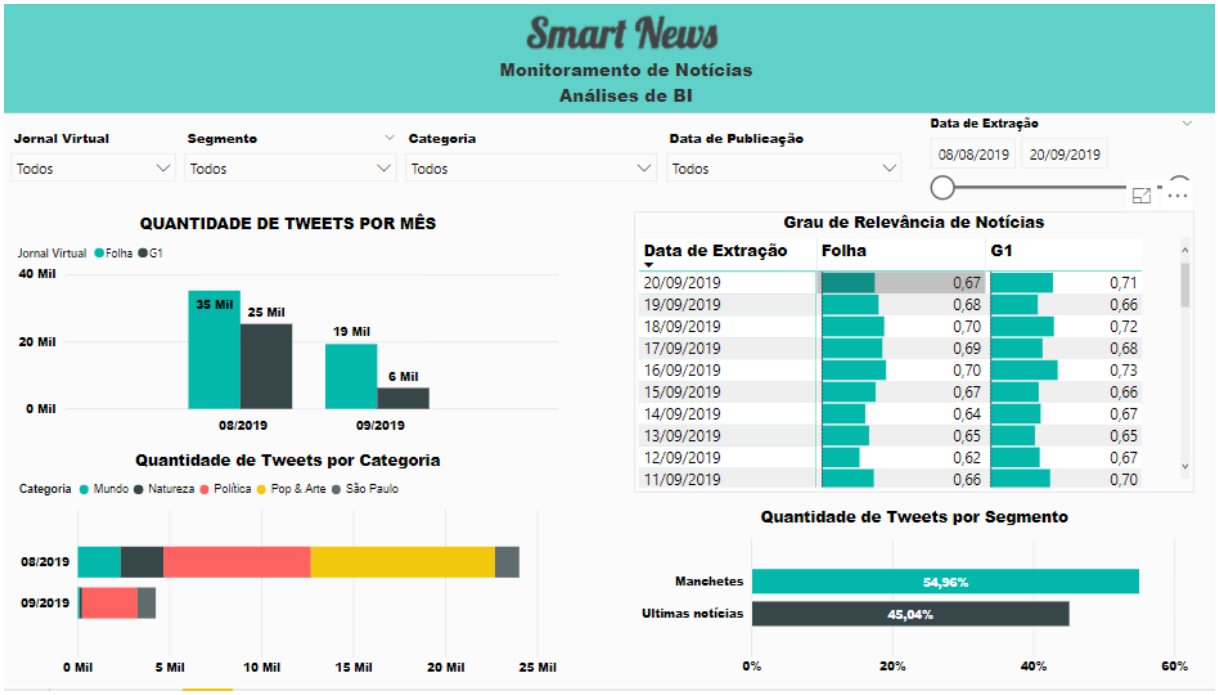
4.13 DASHBOARD

O *dashboard* apresenta-se três tipos diferentes de visualizações, sendo elas: visão geral, visão das notícias e visão das publicações dos *tweets*.

A visão geral apresenta um conjunto de informações gerais do cenário dos dados conforme Figura 4, sendo elas:

- Quantidade de *tweets* do mês por jornal virtual em gráfico de barras;
- Quantidade de *tweets* do mês por categoria em gráfico de barras empilhados;
- Tabela contendo o grau de relevância de notícias por data e jornal virtual;
- Quantidade de *tweets* do mês por categoria em gráfico de barras empilhados;

Figura 4 – Visão Geral



Fonte: a autora

A visão das notícias apresenta um conjunto de informações de cada notícia extraída conforme Figura 5, sendo elas:

- Detalhes da notícia: nome do jornal virtual, data de publicação, segmento e categoria;
- Título da notícia, palavras-chaves, link, frase mais frequente e texto tratado;
- Valores de semelhança da notícia com o mesmo jornal e jornais virtuais diferentes com gráficos de indicadores;

- Valor da relevância da notícia com gráfico de indicador;

Figura 5 – Visão Notícias

Fonte: a autora

A visão das publicações dos *tweets* apresenta um conjunto de informações dos dados *tweets* conforme figura 6, sendo elas:

- Detalhes da *tweet*: texto da publicação, quantidade de palavras e valor referente a análise de sentimento
- Quantidade de *likes* e *retweets* em gráfico de barras;
- Quantidade de publicações por linha de tempo ;
- Valor médio da análise de sentimento dos *tweets*;
- Mapa com localizações dos *tweets*.

Figura 6 – Visão das Publicações dos *Tweets*

Fonte: a autora

4.14 ANÁLISE DO PROTÓTIPO

Os resultados do protótipo proposto neste trabalho podem ser observados nas análises das informações contidas no *dashboard*, o qual conta uma história sobre todo o ciclo de vida de uma notícia e o seu impacto na rede social Twitter.

Assim, obtém-se informações sobre quais fontes de jornais virtuais apresentam maiores números de repercursões por meio de publicações plotadas na linha do tempo, além de quais categorias e segmentos estão sendo mais comentadas na rede social. Com as fontes primárias utilizadas neste protótipo, observa-se que a fonte de jornal Folha apresenta maiores números de publicações e leitores comentando do que a fonte de jornal G1; as categorias com maior número de interações são de políticas e pop e arte, na qual está relacionado com o atual momento do Brasil e perfil de usuários que utilizam a rede social para se interagir. Já os segmentos podem-se notar que as manchetes apresentam melhores resultados que últimas notícias.

A tabela de grau de relevância de notícias, traz o quanto uma é exclusiva em relação a outras. Nota-se que as notícias apresentam um padrão de similaridade entre elas, sendo os resultados muito parecidos entre os dois jornais virtuais selecionados.

Os resultados individuais de cada notícia trazem todos os detalhes que a caracterizam e a identifica, e os dados da etapa de mineração, como as cinco principais palavras-chaves e a frase mais relevânte do texto. Assim, os dados conseguem contribuir para análisar se há correlação entre o título da notícia e o conteúdo principal do texto. Além disso, visualizam-se os valores correspondentes ao grau de similaridade da notícia com outras tanto do mesmo jornal como de outros e o seu grau de relevância. Nota-se que notícias que são publicadas logo após o ocorrido apresentam maior grau de relevância, diminuindo este valor quando outras notícias que abordam o mesmo assunto são publicadas no decorrer do tempo.

Já com os dados dos *tweets* obtêm-se informações do volume de publicações, *likes* e *retweets* referentes a cada notícia para metrificar o real impacto nos leitores por linha do tempo e o comportamento de vida de repercusões dela na rede social.

Nota-se também, o grau de sentimento médio e individual dos leitores para cada publicação, além da quantidade de palavras contidas nas publicações das pessoas para analisar o quanto elas estão falando e como estão enfatizando o assunto.

Para finalizar, têm-se uma visualização de quais regiões as notícias são mais comentadas, nota-se que regiões sudestes apresentam maiores números.

5 CONCLUSÃO

Com o crescimento da digitalização de publicações de matérias e notícias na *web* em tempo real e *online*, onde os jornais estão deixando de seguir os modelos tradicionais de publicações em papeis vendidos em bancas de jornais para estarem em ambientes digitais acessíveis a qualquer pessoa com conexão a internet, há a necessidade de analisar e metrificar o impacto delas nos leitores em ambiente virtual.

O desenvolvimento do protótipo proposto neste trabalho permitiu uma análise para auxiliar o setor jornalistíco a compreender o impacto, relevância e repercussões das notícias na rede social Twitter. Assim, o protótipo possibilita um monitoramento dos jornais virtuais diários e um monitoramento do ambiente de mídia social Twitter de forma integrada e automática, sem precisar de qualquer intervenção humana, ligando as fontes primárias com o ambiente do usuário final, atingido os objetivos traçados.

Para isso, o protótipo fornece um *Data Warehouse* integrado e dimensionado que possibilita consultas em grande escala e manipulação de grande volumes de dados, um portal web de extração e análises de notícias totalmente personalizada para potencializar a autonomia dos usuários do protótipo a solicitar os monitoramentos a qualquer momento quando necessário, agendamento default para realizar os monitoramentos de forma automática, processos de manipulação dos dados baseado nas etapas KDD e disponibilização de informações via *web* por meio de *dashboards* dinâmicos, contidos em um indicador que contém todo o ciclo de vida de uma notícia.

Neste sentido, o protótipo fornece informações sobre comportamentos das notícias e de fontes de informações de maneira personalizada, inteligente e diária para o monitoramento dos resultados no setor jornalístico, auxiliando tomadas de decisões futuras e conhecimento de todo o ciclo de vida de uma notícia, deste da sua publicação, dados e mineração de detalhes até os repercussões e tempo de vida nos leitores.

Além de incentivar e mostrar como a junção do processo de *Business Intelligence* com métodos de Inteligência Artificial são eficientes em análises de comportamento de leitores na rede social.

Como possíveis trabalhos futuros, pode-se inserir métodos de Inteligência Artificial para melhorar o desempenho de cálculo de semelhança de cosseno para análise de similaridade e relevância de textos das notícias, assim algoritmos de redes neurais podem contribuir para encontrar similaridade em palavras semelhantes, considerando o contexto em vez de somente as palavras empregadas nos textos.

REFERÊNCIAS BIBLIOGRÁFICAS

BENEVENUTO, Fabrício; RIBEIRO, Filipe; ARAÚJO, Matheus. Métodos para Análise de Sentimentos em mídias sociais. 2015.

CORMEN, Thomas H.; LEISERSON, Charles Eric; RIVEST, Ronald; RIVEST, Ronald L.; STEIN, Clifford. Algoritmos: Teoria e Prática 6ª. ed.; Rio de Janeiro, Elsevier; 2002.

LOPER, Edward; BIRD, Steven; KLEIN, Ewan. Natural Language Processing with Python; Estados Unidos da América; 2009.

MITCHELL, Ryan. Web Scraping with Python; Estados Unidos da América; 2015.

SANTOS, Maribel Yasmina; RAMOS, Isabel. Business Intelligence : tecnologias da informação na gestão de conhecimento; São Paulo, SP; 2006.

SIDOROV, Grigori; GELBUKH, Alexander; ADORNO, Helena Gomez; PINTO, David. Soft Similarity and Soft Cosine Measure: Similarity of Features in Vector Space Model. Computacion y Sistemas Vol. 18, N, 3, p. 491-504, 2014. Disponível em: <http://www.scielo.org.mx/pdf/cys/v18n3/v18n3a7.pdf>. Acesso em: 05/02/2019.