ROSA HELENA PECCININI SILVA ROSSI

Análise de sentimentos para o auxílio na gestão das cidades inteligentes

São Paulo

2019

ROSA HELENA PECCININI SILVA ROSSI

Análise de sentimentos para o auxílio na gestão das cidades inteligentes

Tese apresentada à Escola Politécnica da
Universidade de São Paulo para obtenção do
Título de Doutora em Ciências

São Paulo

2019

ROSA HELENA PECCININI SILVA ROSSI

Análise de sentimentos para o auxílio na gestão das cidades inteligentes

Tese apresentada à Escola Politécnica da
Universidade de São Paulo para obtenção do
Título de Doutora em Ciências

Área de Concentração: Engenharia de
Computação

Orientador: Prof. Livre-Docente Jorge Rady de
Almeida Junior

São Paulo

2019

Autorizo a reprodução e divulgação total ou parcial deste trabalho, por qualquer meio convencional ou eletrônico, para fins de estudo e pesquisa, desde que citada a fonte.

Este exemplar foi revisado e corrigido em relação à versão original, sob responsabilidade única do autor e com a anuência de seu orientador.

São Paulo, _____ de _____ de _____

Assinatura do autor: _____

Assinatura do orientador: _____

Catálogo-na-publicação

Rossi, Rosa Helena Pecchinini Silva

Análise de sentimentos para o auxílio na gestão das cidades inteligentes
/ R. H. P. S. Rossi -- versão corr. -- São Paulo, 2019.
152 p.

Tese (Doutorado) - Escola Politécnica da Universidade de São Paulo.
Departamento de Engenharia de Computação e Sistemas Digitais.

1.Algoritmos (Classificação) 2.Mineração de dados 3.Cidades inteligentes
4.Computação (Análise-Sentimentos) 5.Cidades inteligentes (Gestão)
I.Universidade de São Paulo. Escola Politécnica. Departamento de
Engenharia de Computação e Sistemas Digitais II.t.

À minha amada Família

AGRADECIMENTOS

A Deus por sempre estar presente na minha vida e por colocar pessoas tão iluminadas nessa jornada, que contribuíram imensamente em aspectos profissionais, acadêmicos e pessoais. A todos, minha gratidão e respeito.

Ao meu orientador, Prof. Jorge Rady de Almeida Junior, pela sua brilhante capacidade de orientação, profissionalismo, comprometimento e, principalmente, pelo seu incentivo, o que me motivou a superar diversos desafios ao longo deste trabalho. Obrigada, Professor! Ser sua orientanda foi uma oportunidade única e muito especial.

Às Professoras Liria Matsumoto Sato e Anarosa Alves Franco Brandão pelas orientações, conhecimentos transmitidos e incentivo recebido durante a qualificação.

Aos funcionários da CET (Compania de Engenharia de Tráfego), extremamente profissionais e colaborativos, e, em especial, ao Ricardo Pires Bastos, por todas as informações recebidas durante a visita técnica na Central de Operações.

À minha mãe e aos meus pais, grandes incentivadores, exemplos de vida, fundamentais na minha formação pessoal. Minha eterna gratidão.

Às minhas filhas Mariana e Juliana, ao meu esposo Raimundo, ao meu genro Fausto, em especial, e ao meu genro Victor. A todos minha gratidão pelo apoio e incentivo em todos os momentos, por me darem força todos os dias e acreditarem no meu trabalho.

Aos meus irmãos Sylvio, Mariana e, em especial, ao meu irmão Alejandro pela ajuda, amizade e apoio de sempre.

Ao meu grande amigo Gilberto Ventura da Luz por ter me motivado a sempre buscar conhecimentos e enfrentar novos desafios, a quem devo também o incentivo pela pesquisa em Análise de Sentimentos.

À bibliotecária Ana Maria Badiali pela amizade, profissionalismo e comprometimento na revisão desta Tese.

Aos colegas Daniel Baraldi Sesso, Ruben Glatt, Lucio F. Vismari, Desiré Nguessan, Hellinton Hatsuo, José Luiz Pagnossim, Ignacio Sánchez Gendriz pela troca de conhecimentos técnicos e pela grande amizade, essenciais para a conclusão deste trabalho.

A todos os funcionários da Poli que contribuíram direta ou indiretamente para que esta Tese se concretizasse.

“Apenas dê o primeiro passo, mesmo sem saber o quão desafiador e longo será o caminho. Nunca desista. Tenha fé em Deus durante todo o percurso de sua vida. Com certeza, pessoas iluminadas surgirão para te acompanhar.”

RESUMO

Esta Tese tem como objetivo geral inserir a Análise de Sentimentos na gestão das Cidades Inteligentes, possibilitando a implementação de uma ferramenta que disponibilize informações que auxiliem na supervisão e gestão dessas cidades. Dentre os possíveis auxílios que podem ser prestados está a identificação de ações, meios de prevenção e predição de possíveis adversidades nos diversos Domínios de Interesse, além da busca por melhorias na qualidade de vida da população, que pode ser feita por meio dessa análise, permitindo que os gestores dessas cidades possam tomar as melhores decisões de acordo com cada cenário. Este trabalho contribui com um novo método cujo o objetivo é o desenvolvimento de um Sistema de Análise de Sentimentos para Auxílio na Gestão das Cidades Inteligentes (ASCI). Esse Sistema é capaz de captar, tratar, processar, filtrar por Domínio de Interesse e avaliar os sentimentos contidos nas informações provenientes dos cidadãos de uma Cidade Inteligente. O método utiliza duas Fases de Mineração de Dados, uma para a classificação dos Domínios de Interesse e outra para a Análise de Sentimentos. Para o estudo de caso foi implementado o método ASCI por meio do qual são captadas informações provenientes da população de uma determinada região da cidade de São Paulo, por meio da Rede Social Twitter. Também foi realizado um estudo de classificação de sentimentos no Domínio específico do Transporte, no qual também foram utilizados, e tiveram seu desempenho avaliado, os classificadores do tipo Linear SVC, Logistic Regression, Multinomial Naive Bayes e Random Forest Classifier para identificar os sentimentos positivos, neutros e negativos dos tweets captados. Os dados foram avaliados usando duas técnicas de extração de características de texto: Bag of Words e TF-IDF. O método ASCI desenvolvido nesta Tese contribui de maneira relevante para a área de Análise de Sentimentos, uma vez que os resultados obtidos foram satisfatórios quando aplicado em cenários de Domínios de Interesse das Cidades Inteligentes.

Palavras-Chave: Cidades Inteligentes, Análise de Sentimentos, Mineração de Dados, Algoritmos de Classificação.

ABSTRACT

The main objective of this work is to insert the Sentiment Analysis in the management of Smart Cities, enabling the implementation of a supervision and management tool in these cities. Among the possible aid services that can be applied, there is the identification of actions, ways of prevention and prediction of possible adversities in the various Domains of Interest, and also the search for improvements in the quality of life of the population. This can be done through this analysis, allowing the best decisions according to each scenario by the city managers. This work contributes to a new method whose objective is the development of a Sentiment Analysis System to Assist in the Management of Smart Cities (ASCI). This System is capable of capturing, classifying, processing, filtering by Domain of Interest and evaluating the sentiments of Smart City citizens. The method uses two Data Mining phases, one for the classification of Domains of Interest and the other for Sentiment Analysis. For the case study, the ASCI method was implemented, through which information was collected from a regional population in São Paulo city through Twitter Social Network data. A study of Sentiment Analysis in specific Domain of Interest Transport was also carried out, in which Linear SVC, Logistic Regression, Multinomial Naive Bayes and Random Forest classifiers were used to identify the positive, neutral and negative sentiments of collected tweets. The data were evaluated using two techniques of extraction of text characteristics: Bag of Words and TF-IDF. The ASCI method developed in this Thesis contributes significantly to the area of Sentiment Analysis and the results obtained were satisfactory when applied in Smart City Domain of Interest scenarios.

Keywords : Smart Cities, Sentiment Analysis, Data Mining, Classification Algorithms.

LISTA DE FIGURAS

Figura 1	- Etapas do processo KDD	16
Figura 2	- Árvore de Decisão	19
Figura 3	- Processo para extrair Insights do Big Data	26
Figura 4	- Implantação de Sensores e Geração de Dados do Sistema Inteligente	34
Figura 5	- Elementos-chave das Cidades Inteligentes e Humanas	36
Figura 6	- Categorias de Sentenças	44
Figura 7	- Níveis de Análise de Sentimentos	47
Figura 8	- Léxico de Sentimentos	50
Figura 9	- Fases do Desenvolvimento do Sistema de Análise de Sentimentos para Auxílio na Gestão das Cidades Inteligentes	58
Figura 10	- Processo de Preparação da Representação dos Dados.....	60
Figura 11	- Bag of Words Matriz de Representação das Ocorrências das Palavras	62
Figura 12	- Método de Desenvolvimento do Sistema de Análise de Sentimentos para Auxílio na Gestão das Cidades Inteligentes.....	67
Figura 13	- Matriz de Predição	68
Figura 14	- Matriz de Predição Multiclasse	70
Figura 15	- Mapa do Município de São Paulo subdividido por área de atuação das GETs	72
Figura 16	- Central de Operações - CET	74
Figura 17	- Gráfico de Lentidão na Cidade de São Paulo.....	75
Figura 18	- Mapa de Lentidão na Cidade de São Paulo.....	75
Figura 19	- Mapa do Centro da Cidade de São Paulo	78
Figura 20	- Remoção de Pontuações, Números, Caracteres Especiais e Palavras com até 2 Letras	80
Figura 21	- Tokenização e Stemming dos textos	80

Figura 22 - Nuvem de Palavras referente ao domínio de Transporte	81
Figura 23 - Nuvem de Palavras referente ao domínio da Segurança	81
Figura 24 - Nuvem de Palavras referente ao domínio da Educação	82
Figura 25 - Distribuição do Corpus de Tweets para Treinamento e Teste	83
Figura 26 - Gráfico de Distribuição de Todo conjunto de Tweets	84
Figura 27 - Validação Cruzada - Método K-fold	85
Figura 28 - Boxplot da Acurácia da Classificação nos três Domínios – Técnica BOW ..	85
Figura 29 - Boxplot da Acurácia da Classificação nos três Domínios – Técnica TF-IDF	86
Figura 30 - Desempenho do algoritmo LinearSVC – Treinamento nos Domínios	87
Figura 31 - Desempenho do algoritmo Logistic Regression - Treinamento nos Domínios	87
Figura 32 - Desempenho do algoritmo Multinomial Naive Bayes - Treinamento nos Domínios	88
Figura 33 - Desempenho do algoritmo Random Forest - Treinamento nos Domínios	89
Figura 34 - Validação do Treinamento – Classificação dos Tweets por Domínios	90
Figura 35 - Gráficos Validação do Treinamento – Classificação dos Tweets por Domínio	90
Figura 36 - Desempenho do classificador Logistic Regression nos Domínios - Corpus Teste	91
Figura 37 - Classificação dos Tweets por Domínios – Corpus Teste	91
Figura 38 - Gráficos Classificação dos Tweets por Domínios – Corpus Teste	92
Figura 39 - Tweets por Domínios - Corpus Teste	92
Figura 40 - Principais Hashtags dos Tweets do Corpus Teste – Domínios	93
Figura 41 - Gráfico de Distribuição do conjunto de Tweets, Treinamento e Teste, do Domínio Transporte	94
Figura 42 - Boxplot da Acurácia da Classificação dos Sentimentos no Domínio Transporte – Técnica BOW	94

Figura 43 - Boxplot da Acurácia da Classificação dos Sentimentos no Domínio Transporte – Técnica TF-IDF	95
Figura 44 - Desempenho do algoritmo LinearSVC – Treinamento dos Sentimentos no Domínio Transporte	96
Figura 45 - Desempenho do algoritmo Logistic Regression - Treinamento dos Sentimentos no Domínio Transporte	97
Figura 46 - Desempenho do algoritmo Multinomial Naive Bayes - Treinamento dos Sentimentos no Domínio Transporte	97
Figura 47 - Desempenho do algoritmo Random Forest - Treinamento dos Sentimentos no Domínio Transporte	98
Figura 48 - Validação do Treinamento – Classificação dos Tweets por tipo de Sentimento	99
Figura 49 - Gráficos Validação doTreinamento – Classificação dos Tweets por tipo de Sentimento	99
Figura 50 - Desempenho do classificador LinearSVC – Sentimentos no Domínio Transporte - Corpus Teste	100
Figura 51 - Classificação dos Tweets por Sentimentos no Domínio Transporte - Corpus Teste	100
Figura 52 - Gráficos Classificação dos Tweets por Tipo de Sentimento no Domínio Transporte - Corpus Teste	101
Figura 53 - Nuvem de Palavras com Sentimentos Positivo das Informações dos Tweets no Domínio do Transporte	101
Figura 54 - Nuvem de Palavras com Sentimentos Negativo das Informações dos Tweets no Domínio do Transporte	102
Figura 55 - Nuvem de Palavras com Sentimentos Neutro das Informações dos Tweets no Domínio do Transporte	102
Figura 56 - Tweets por Tipo de Sentimento no Domínio do Transporte – Corpus Teste	103
Figura 57 - Principais Hashtags dos Tweets do Corpus Teste – Sentimento	103

LISTA DE TABELAS

Tabela 1 - Exemplos de frases e suas respectivas polaridades	43
Tabela 2 - Resumo dos Principais Trabalhos em Análise de Sentimento Multiclasse	55
Tabela 3 - Exemplo de Caracteres Especiais	60
Tabela 4 - Exemplos de Unigrama e Bigrama	62
Tabela 5 - Tabela Resumo de Desempenho dos Classificadores por Domínio	104
Tabela 6 - Tabela de Avaliação do Melhor Classificador por Domínio com o Corpus de Teste	105
Tabela 7 - Tabela Resumo de Desempenho dos Classificadores para Análise de Sentimentos	105
Tabela 8 - Tabela de Avaliação do Melhor Classificador para Análise de Sentimentos	106
...	

LISTA DE ABREVIATURAS

API – Application Programming Interface

ASCI – Sistema de Análise de Sentimentos para o Auxílio na Gestão das Cidades Inteligentes

BOW – Bag of Words

CET – Companhia de Engenharia de Tráfego

DT – Decision Tree

IEEE – Institute of Electrical and Electronics Engineers

IoT – Internet of Things

MCS – Mobile Crowd Sensing

ML – Machine Learning

NB – Naive Bayes

NN – Neural Network

OSN – Online Social Networks OSN

PLN – Processamento de Linguagem Natural

SC – Smart Cities

SHC – Smart Human Cities

SVM – Support Vector Machine

TF-IDF – Term Frequency – Inverse Document Frequency

TIC – Tecnologias da Informação e Comunicação

SUMÁRIO

1. INTRODUÇÃO	10
1.1. Contextualização	10
1.2. Motivação e justificativa	11
1.3. Objetivos	12
1.4. Metodologia	12
1.5. Organização do trabalho	14
2. MINERAÇÃO DE DADOS	15
2.1. Conceitos de Mineração de Dados	15
2.2. Funções de Mineração de Dados	17
2.2.1. Classificação	17
2.2.2. Agrupamento (Clustering)	21
2.2.3. Análise de Associação	21
2.2.4. Análise de Séries Temporais	22
2.2.5. Análise de Outliers	22
2.3. Big Data	23
2.3.1. Principais Ferramentas de Big Data	23
2.3.2. Limpeza dos Dados	25
2.3.3. Análise e Gerenciamento de Dados	25
2.4. Aplicações de Mineração de Dados na Governança das Cidades	27
3. REDES SOCIAIS	28
3.1. Cenário das Redes Sociais no Mundo	29
3.2. Redes Sociais e Cidades Inteligentes	30
4. CIDADES INTELIGENTES (SMART CITIES)	32
4.1. Internet das Coisas (IoT), Big Data e as Cidades Inteligentes	33
4.2. Smart Human Cities (SHC)	35
4.3. Estudos de Casos das Cidades Inteligentes	37
5. ANÁLISE DE SENTIMENTOS	40
5.1. Termos e Conceitos Utilizados em Análise de Sentimentos	42
5.1.1. Definições	42
5.1.2. Polaridade	43
5.1.3. Categorização das Sentenças: Subjetividade x Objetividade	44
5.1.4. Tipos de Opinião ou Análise de Sentimentos	45
5.1.5. Semântica	46

5.1.6.	Lidando com Figuras de Linguagem	46
5.2.	Níveis de Análise de Sentimentos	47
5.2.1.	Nível do Documento (Mensagem)	47
5.2.2.	Nível de Sentença	48
5.2.3.	Nível de Aspecto	48
5.2.4.	Nível de Palavra ou Dicionário	48
5.3.	Métodos de Classificação para Análise de Sentimentos	49
5.3.1.	Classificadores Baseados em Léxico	49
5.3.2.	Classificadores baseados em Aprendizagem de Máquina Supervisionada	51
5.4.	Trabalhos Correlatos	53
5.5.	Considerações Finais do Capítulo	55
6.	PROPOSTA DO MÉTODO DE DESENVOLVIMENTO DO SISTEMA DE ANÁLISE DE SENTIMENTOS PARA O AUXÍLIO NA GESTÃO DAS CIDADES INTELIGENTES (ASCI)	57
6.1.	Definição do Método de Desenvolvimento do Sistema de Análise de Sentimentos para Auxílio na Gestão das Cidades Inteligentes	57
6.1.1.	1ª Fase - Captação dos Dados	58
6.1.2.	2ª Fase - Pré-processamento dos Dados	59
6.1.3.	3ª Fase - Mineração de Dados - Classificação dos Domínios	63
6.1.4.	4ª Fase - Mineração de Dados - Análise de Sentimentos	65
6.1.5.	Fluxo Geral das Informações do ASCI	66
6.2.	Métricas para Avaliação do Desempenho do Classificador	68
6.3.	Uso do ASCI na Gestão das Cidades Inteligentes	70
7.	ESTUDO DE CASO - SISTEMA DE GESTÃO DAS CIDADES INTELIGENTES UTILIZANDO ANÁLISE DE SENTIMENTOS	71
7.1.	Estudo de Caso: CET - Companhia de Engenharia de Tráfego	71
7.1.1.	Central de Operações	73
7.1.2.	Ocorrências no Trânsito	76
7.2.	Implementação do Estudo de caso Sistema de Análise de Sentimento para Auxílio na Gestão das Cidades Inteligentes - ASCI	76
7.2.1.	1ª Fase realização da Captação dos Dados	77
7.2.2.	2ª Fase: realização do Pré-processamento dos Dados	79
7.2.3.	3ª Fase - Mineração de Dados, Classificação dos Domínios Transporte, Segurança e Educação	83
7.2.4.	4ª Fase - Mineração de Dados - Análise de Sentimentos	93
7.3.	Considerações Finais do Capítulo	104
8.	CONCLUSÕES E TRABALHOS FUTUROS	107
8.1.	Conclusões	107
8.2.	Trabalhos Futuros	108
	REFERÊNCIAS	110

ANEXO I - PROGRAMA DE CAPTAÇÃO DE TWEETS ATRAVÉS DA API DO TWITTER	115
---	------------

ANEXO II - PROGRAMAS DE IMPLEMENTAÇÃO DA METODOLOGIA E ESTUDO DE CASO	118
--	------------

1. INTRODUÇÃO

Neste capítulo são apresentadas a contextualização do trabalho, motivações justificativas e objetivo, além da metodologia de trabalho.

1.1. Contextualização

O êxodo rural é uma realidade da sociedade pós-moderna e globalizada. O movimento de migração do campo para a cidade foi influenciado pela mecanização da produção agropecuária e pela busca de uma melhor qualidade de vida.

Desde 2009, a maior parte da população mundial passou a viver nos centros urbanos das cidades. O contingente de pessoas morando em zonas urbanas já representava mais de 51% da população mundial (ONUBR - NAÇÕES UNIDAS NO BRASIL, 2016). Em 2011, de acordo com o crescimento das cidades, há uma preocupação dos governantes em elaborar sistemas complexos de infraestrutura e recursos que atendam à população (CARAGLIU; BO; NIJKAMP, 2011). Atualmente, 55% da população mundial vive em áreas urbanas e é previsto que até o ano de 2050 este percentual aumente para 68% da população mundial (ONU, 2019).

Esse fato acrescenta desafios às autoridades governamentais referentes à gestão de problemas relacionados a temas específicos, como: abastecimento de água, sistema energético, eliminação de resíduos locais, sistema de gestão do tráfego urbano, saúde, educação, economia, meio ambiente, turismo, segurança pública e outros.

Em paralelo ao crescimento urbano, houve também um desenvolvimento tecnológico que possibilitou o desenvolvimento do ambiente da Internet das Coisas (IoT), no qual há um grande número de objetos inteligentes conectados à Rede, monitorando, processando dados e permitindo a comunicação entre objetos e usuários, objetos e objetos, e usuários e usuários, evoluindo para uma comunicação cada vez mais ubíqua (ZANELLA et al., 2014).

A Internet é um ambiente com uma infinidade de possibilidades no qual as pessoas podem buscar e compartilhar informações, expressar opiniões e sentimentos a respeito de qualquer assunto, a qualquer hora e local, e na atualidade, principalmente por meio de redes sociais (Facebook, Twitter, Google, WhatsApp, etc), sites, blogs e outros. Dessa forma, uma enorme quantidade de dados é gerada e disponibilizada na Web a cada dia

De acordo com Lewis (2018), citam-se alguns dados de volume de informações que circulam na Internet em 60 de segundos: 973 mil logins no Facebook, 481 mil tweets enviados, 38 milhões de mensagens pelo WhatsApp, 187 milhões de e-mails enviados, dentre outros.

Todos esses cenários oferecem as condições ideais para o desenvolvimento das Smart Cities (Cidades Inteligentes), conforme previsto pelo IEEE, onde uma cidade inteligente congrega tecnologia, governo e sociedade para viabilizar algumas características: inteligência em economia, inteligência em mobilidade, inteligência em meio ambiente, inteligência em pessoas, inteligência em qualidade de vida e inteligência em governança (IEEE Smart Cities, 2017).

Alguns pesquisadores dessa área também reconhecem que o uso das informações provenientes dos cidadãos, através da análise de sentimentos ou mineração de opinião, pode contribuir para uma melhor compreensão e elaboração de respostas por parte dos governos das cidades de acordo com as necessidades e preocupações dos habitantes. No entanto, relativamente pouco se sabe sobre os benefícios potenciais da Análise de Sentimentos em proveito das Cidades Inteligentes (AHMED et al., 2016).

1.2. Motivação e justificativa

O advento das Cidades Inteligentes, da Internet das Coisas e das plataformas de mídia social transformaram a maneira como as pessoas se comunicam e se expressam. Diariamente, novos dispositivos inteligentes são disponibilizados para a sociedade. As pessoas estão compartilhando suas experiências e pontos de vista por meio de postagens e comentários em blogs, sites e redes sociais utilizando textos, fotos e vídeos. Reconheceu-se que as reações oportunas às opiniões e sentimentos públicos e seu uso adequado pelos governos municipais são de suma importância para suas tomadas de decisão. (AHMED; BOUHORMA, 2016).

Há um grande interesse em extrair insights valiosos dos grandes volumes de conteúdo gerados pelos usuários diariamente. Transformar esses dados em informações valiosas pode ser significativo e útil para a governança das cidades ou até mesmo para cidadãos comuns que desejam ser constantemente informados sobre serviços de utilidade pública de suas cidades.

A motivação deste trabalho baseia-se na oportunidade de contribuição da Análise de Sentimentos da população para auxílio na gestão das Cidades Inteligentes no ambiente da IoT, visando, principalmente, auxiliar na análise, avaliação e prevenção de situações de risco nas mais diversas áreas, de forma que os gestores e órgãos governamentais possam tomar as melhores decisões, utilizando os recursos necessários de acordo com cada cenário.

1.3. Objetivos

Esta Tese tem como objetivo geral inserir a Análise de Sentimentos na gestão das chamadas Cidades Inteligentes, incluindo, dessa forma, mais uma ferramenta de supervisão e controle dessas cidades. Dentre os possíveis auxílios que podem ser prestados está a identificação de ações, meios de prevenção e predição de possíveis adversidades além da busca por melhorias na qualidade de vida da população, o que pode ser feito por meio dessa análise, permitindo que os gestores dessas cidades possam tomar melhores decisões de acordo com cada cenário.

Como objetivo específico, esta pesquisa visa o concebimento de um Método de Desenvolvimento de um Sistema de Análise de Sentimentos para Auxílio na Gestão das Cidades Inteligentes (ASCI), realizando a captação das informações disponibilizadas por cidadãos das Cidades Inteligentes, nas Redes Sociais. O método é baseado nas etapas do processo Knowledge Discovery in Databases (KDD), com a inserção de uma etapa a mais de mineração de dados. A primeira etapa de mineração de dados consiste em classificar o domínio de interesse da Gestão da Cidade Inteligente, por exemplo: Transporte, Segurança, Educação e outros. Já a segunda etapa de mineração de dados é dedicada à Análise de Sentimentos no domínio anteriormente classificado.

Como premissa deste trabalho, considera-se que os cidadãos da Cidade Inteligente estejam conectados à Internet, sendo então possível realizar a captação de informações desses municípios por meio das Redes Sociais, além de se considerar também o ambiente da IoT, que disponibiliza informações que podem ser agregadas aos dados captados nas Redes Sociais.

1.4. Metodologia

A Metodologia aplicada neste trabalho compreendeu as seguintes atividades:

- Levantamento Bibliográfico: foram estudados assuntos como Mineração de Dados, Cidades Inteligentes e Análise de Sentimentos.

A maioria dos artigos pesquisados foi proveniente do portal de periódicos da CAPES (<http://www.periodicos.capes.gov.br/>) e dos seguintes bancos de dados de artigos e conferências:

- ACM Digital Library (<http://dl.acm.org/>)
 - Elsevier (<http://www.elsevier.com/>)
 - IEEE Xplore Digital Library (<http://ieeexplore.ieee.org/>)
 - Scopus (<http://www.scopus.com/>)
 - Science Direct (<http://www.sciencedirect.com/>)
 - Springer Link (<http://link.springer.com/>)
- Estudo de técnicas de Análise de Sentimentos: avaliação das principais tecnologias usadas em Processamento de Linguagem Natural para Análise de Sentimento;
 - Definição da técnica utilizada neste trabalho, no caso Aprendizado de Máquina Supervisionado com a utilização de Bag of Words (BOW) e TF-IDF (Term Frequency – Inverse Document Frequency);
 - Definição dos algoritmos de classificação do tipo Support Vector Machine, Logistic Regression, Decision Tree e Naive Bayes;
 - Elaboração do Método de Desenvolvimento de um Sistema de Gestão das Cidades Inteligentes com o uso de Análise de Sentimentos;
 - Elaboração de um estudo de caso utilizando o método acima descrito.

1.5. Organização do trabalho

O Capítulo 1 apresenta uma seção introdutória, o contexto, a motivação, a justificativa e o objetivo do trabalho.

O Capítulo 2 apresenta os principais conceitos de Mineração de Dados.

O Capítulo 3 apresenta uma revisão sobre o tema Redes Sociais.

O Capítulo 3 apresenta o estado da arte referente ao tema Cidades Inteligentes.

O Capítulo 4 constitui uma revisão teórica dos conceitos básicos envolvidos em Análise de Sentimentos, desde sua definição, níveis e métodos de análise.

O Capítulo 5 constitui uma revisão teórica dos conceitos básicos de Redes Sociais.

O Capítulo 6 apresenta um Método de Desenvolvimento de um Sistema de Análise de Sentimentos para Auxílio na Gestão das Cidades Inteligentes (ASCI)

O Capítulo 7 apresenta um Estudo de Caso usando o Método ASCI na área de Transporte de uma determinada região da cidade de São Paulo.

O Capítulo 8 apresenta as considerações e conclusões finais e também são descritos os trabalhos futuros relacionados ao tema.

Por último são apresentadas as Refências utilizadas e, na sequência, os Anexos com os códigos fontes deste trabalho.

2. MINERAÇÃO DE DADOS

A sociedade atual está gerando e coletando informações em uma escala nunca antes vista na história da humanidade. Computadores pessoais facilitam salvar informações que antigamente seriam destruídas e com a possibilidade do armazenamento de dados cada vez mais barata, torna-se muito mais fácil adiar decisões sobre o que fazer com essas informações, simplesmente tudo é guardado. Além disso, a tecnologia registra todas as decisões dos usuários, inclusive suas escolhas. Por exemplo, os produtos que são comprados no supermercado, hábitos financeiros, lugares que visitados com mais frequência, ou seja, a maior parte das ações comuns do dia a dia são registradas (WITTEN; EIBE, 2005).

Na Web, bilhões de pesquisas são realizadas diariamente suportadas por mecanismos de busca. Comunidades e mídias sociais também são fontes de dados cada vez mais importantes, pois nelas os usuários disponibilizam e compartilham imagens, vídeos, textos e outros conteúdos. Essa enorme quantidade de dados, que é coletada e armazenada em grandes e numerosos repositórios de dados, excede em muito nossas habilidades humanas de análise e compreensão. Como resultado, uma gigantesca quantidade de dados é raramente visitada e a falta de ferramentas adequadas de análise de dados impede que todo esse conteúdo seja tratado e analisado de maneira apropriada. Esse fato tem sido descrito como uma situação rica em dados, mas pobre em informação (HAN, JIAWEI; KAMBER, MICHELINE; PEI, 2012).

2.1. Conceitos de Mineração de Dados

A Mineração de Dados envolve a descoberta de padrões novos e potencialmente úteis em grandes conjuntos de dados e também a aplicação de algoritmos para a extração de informações ocultas. Todo esse processo é importante para adquirir conhecimento a partir dos dados e, posteriormente, para as tomadas de decisões. O objetivo de qualquer processo de mineração de dados é construir um modelo eficiente, descritivo ou preditivo, de uma grande quantidade de dados que não apenas melhor se adapte ou explique, mas que também seja capaz de se generalizar a novos dados (CHEN et al., 2015).

Historicamente, outros termos têm sido usados para Mineração de Dados (Data Mining), por exemplo: Descoberta de Conhecimento em Bancos de Dados (Knowledge Discovery in Databases - KDD), Extração de Conhecimento, Análise de Dados/ Padrões, Arqueologia de

Dados, Dragagem de Dados e Coleta de Informações (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

Para alguns autores, eles são considerados sinônimos e para outros, como Fayyad et al. (1996), o KDD refere-se a todo o processo de descoberta de conhecimento, e a Mineração de Dados é uma das etapas desse processo. No entanto, todos concordam que o processo de mineração deve ser iterativo, iterativo e dividido em fases. A Figura 1 apresenta o processo de KDD (CAMILO; SILVA, 2009).

Figura 1 - Etapas do processo KDD

Fonte: Autora adaptada de (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996)

O processo de KDD envolve várias etapas: selecionar os dados de destino, pré-processar os dados, transformando-os, se necessário, realizar a Mineração de Dados para extrair padrões e relacionamentos e, em seguida, interpretar e avaliar as avaliações das estruturas descobertas (HAND, DAVID; MANNILA; SMYTH, 2001).

2.2. Funções de Mineração de Dados

As funções de Mineração de Dados (Data Mining) incluem: Classificação, Agrupamento, Análise de Associação, Análise de Séries Temporais e Análise de Outlier (CHEN et al., 2015).

2.2.1. Classificação

A classificação é importante para o gerenciamento de tomada de decisão. O processo chamado de classificação é a atribuição de um objeto a uma das categorias ou classes de destino predefinidas, ou seja, prever com precisão a classe alvo para cada caso de dados. Por exemplo, um modelo de classificação poderia ser usado para identificar candidatos a empréstimos classificando-os como baixo, médio ou alto risco de crédito.

Existem vários métodos para classificar os dados, tais como: Máquinas de Vetores de Suporte SVM (Support Vector Machine), Árvore de Decisão (Decision tree), Algoritmo Vizinho mais Próximo KNN (K-Nearest Neighbor), Redes Bayesianas (Bayesian Network) e, a seguir descritos:

- O algoritmo Máquinas de Vetores de Suporte (Support Vector Machines -SVM) é um modelo de aprendizado supervisionado, que analisa dados e reconhece padrões. É um classificador binário que forma um hiperplano e separando o espaço de recurso de instâncias positivas do espaço de recurso de instâncias negativas. Durante a fase de treinamento, o hiperplano separa o espaço de atributos positivo do espaço de atributos negativo com uma margem máxima escolhida. A margem é a distância do ponto mais próximo dos conjuntos positivo e negativo para o hiperplano. Os vetores de suporte representam o subconjunto das instâncias de treinamento, determinam o hiperplano para um SVM. No caso de multiclases vários hiperplanos são gerados (CHEN et al., 2015);

- Regressão Logística (Logistic Regression): É usada quando o resultado do interesse é dicotômico (ou policotômico) por natureza e os preditores de interesse podem ser categóricos ou contínuos. ((DAVIS; OFFORD, 1997). A regressão logística usa um modelo denominado logit, sendo que logit está diretamente relacionado à probabilidade do resultado. A relação entre o logit e a probabilidade pode ser visualizada por meio da Equação 1 (SAINANI, 2014):

$$\text{logit} = \ln (p/1-p) \quad (1)$$

- Árvore de Decisão (Decision Tree - DT): um classificador de Árvores de Decisão assemelha-se a uma árvore na qual os recursos são representados por nós, as arestas que deixam um nó são rotuladas pelo peso do recurso e as folhas representam as categorias. A árvore é construída com base em um procedimento recursivo. A cada passo é escolhida uma característica 'F' e a coleção de treino é dividida em dois grupos, um contendo 'F' e outro não contendo 'F'. Este procedimento é realizado até que permaneçam apenas documentos de uma única categoria. Uma folha é gerada no final deste procedimento. Ganho de informação ou entropia é usado para escolher um recurso em cada etapa. Por exemplo, método de classificação que, constrói uma figura em forma de árvore, onde cada nó (não folha) indica um teste feito sobre um valor (por exemplo, idade > 20). As ligações entre os nós representam os valores possíveis do teste do nó superior, e as folhas indicam a classe (categoria) a qual o registro pertence. Um exemplo de uma árvore de decisão é apresentado na Figura 2;

Figura 2 - Árvore de Decisão

Fonte: Autora

- O algoritmo K Vizinhos mais Próximos (K Nearest Neighbors – KNN) é empregado principalmente na análise de prognósticos. Ele interpreta dados como um ponto no espaço, definindo o conceito de vizinhança, ou seja, identifica o conjunto de registros que estão próximos por alguma característica. Ao estimar a renda de um indivíduo de uma população, pesquisando $k=20$ vizinhos mais próximos, através das informações de valores dos atributos bairro de moradia, profissão, escolaridade e idade. Para o uso desta técnica é necessário existir registros com um número de atributos suficientes para determinação da vizinhança (CÔRTEZ; PORCARO; LIFSCHITZ, 2002);
- Classificação Bayesiana (Naive Bayes – NB) é uma técnica estatística baseada no teorema de Thomas Bayes, com a qual é possível encontrar a probabilidade de um certo evento ocorrer, dada a probabilidade de um outro evento que já ocorreu: Os algoritmos Bayesianos, também chamados de naive Bayes, partem do princípio que não existe relação de dependência entre os atributos. No entanto, quando não for possível, uma variação conhecida como Bayesian Belief Networks, ou Bayesian Networks, deve ser utilizada (CAMILO; SILVA, 2009).

Os classificadores probabilísticos usam o teorema de Bayes para calcular a probabilidade $P(c | d)$, que um documento pertença a uma determinada categoria c . Conforme a Equação 2:

$$P(c/d) = \frac{P(d/c) * P(c)}{P(d)} \quad (2)$$

$P(c / d)$ é a probabilidade posterior da classe (c , alvo) dada preditor (d , atributos).

$P(c)$ é a probabilidade original da classe.

$P(d / c)$ é a probabilidade que representa a probabilidade de preditor dada a classe.

$P(d)$ é a probabilidade original do preditor.

Este algoritmo é um classificador probabilístico que calcula um conjunto de probabilidades a partir do cálculo de frequência e combinações dado um conjunto de dados. O algoritmo utiliza o teorema de Bayes e assume que todos atributos são independentes dado o valor da classe. Dado um conjunto de dados rotulados, a etapa de treinamento consiste em calcular a probabilidade de ocorrência cada atributo em relação às classes. Na etapa de teste, o algoritmo calcula a probabilidade de cada atributo de um exemplo não visto ocorrer com base nas probabilidades obtida na etapa de treino. O exemplo não visto é classificado com a classe que possui maior probabilidade de ocorrer. Esse algoritmo é considerado ingênuo, uma vez que ele considera que todos os atributos são independentes durante o cálculo dessas probabilidades (DOMINGOS; PAZZANI, 1997);

- **Redes Neurais:** As Redes Neurais (Neural Networks - NN) podem ser projetadas para realizar tarefas de mineração de opinião. Os recursos de um documento são os nós de entrada, os nós de saída entregam a categoria. As relações de dependência são atendidas pelos pesos de ligação. Geralmente, as NN são treinadas por retrotransmissão, isto é, os documentos alimentados nos nós de entrada e se ocorrer uma classificação errada, o erro é propagado de volta na rede para minimizar o erro, ajustando os pesos de ligação. Perceptron é o tipo mais simples de uma NN, que tem

apenas duas camadas: camada de entrada e camada de saída. Um perceptron de várias camadas contém uma ou mais camadas ocultas entre as camadas de entrada e de saída.

2.2.2. Agrupamento (Clustering)

A tarefa de agrupamento visa identificar e aproximar os registros similares. Um agrupamento (ou cluster) é uma coleção de registros similares entre si, porém, diferente dos outros registros nos demais agrupamentos, ela apenas identifica os grupos de dados similares. Por exemplo, um modelo de agrupamento poderia identificar a segmentação de mercado, reconhecer padrões, processar imagens, analisar dados, fazer a taxonomia de plantas e animais, detectar fraudes, entre outras (CAMILO; SILVA, 2009).

Algumas técnicas para agrupamento são (CAMILO; SILVA, 2009):

- **Métodos Hierárquicos:** a ideia básica dos métodos hierárquicos é criar o agrupamento por meio da aglomeração ou da divisão dos elementos do conjunto. A forma gerada por estes métodos é um dendrograma;
- **Métodos de Particionamento:** dado um conjunto D de dados com n registros e k o número de agrupamentos desejados, os algoritmos de particionamento organizam os objetos em k agrupamentos, tal que $k \leq n$. Os algoritmos mais comuns de agrupamento são k-Means e k-Medoids;
- **Métodos Baseados na Densidade:** os métodos de particionamento e hierárquicos geram agrupamentos de formato esféricos (distribuição dos valores dos dados é mais esparsa). No entanto, existem situações em que essa distribuição é mais densa e que tais métodos não apresentam resultados satisfatórios. Os métodos baseados na densidade conseguem melhores resultados.

2.2.3. Análise de Associação

A mineração de Regras de Associação se concentra na análise de dados de transação. Dado um conjunto de transações, ela busca encontrar regras que prevejam a ocorrência de um item baseado nas ocorrências de outros itens da transação.

Alguns exemplos que podem ser citados são: determinar os casos onde um novo medicamento pode apresentar efeitos colaterais, identificar os usuários que respondem bem a uma oferta de novos serviços, identificar num supermercado itens de compra que tendem a ser vendidos associadamente a outros itens em momentos específicos (CHEN et al., 2015).

Este método consiste em identificar o relacionamento dos itens mais frequentes em um determinado conjunto de dados, por exemplo: SE compra LEITE e PÃO TAMBÉM compra MANTEIGA (CAMILO; SILVA, 2009).

2.2.4. Análise de Séries Temporais

A análise de séries temporais inclui métodos e técnicas para analisar dados de séries temporais, a fim de extrair estatísticas significativas e outras características dos dados (CHEN et al., 2015).

Por exemplo, bases que armazenam o valor das ações de um mercado, velocidade do vento, medidas da atmosfera. O processo de identificação de padrões em bases desse tipo envolve outras técnicas e análises. Para a detecção de fatores de risco na área médica, é utilizada a mineração de séries temporais por meio de algoritmos de agrupamento.

2.2.5. Análise de Outliers

Descreve e modela regularidades ou tendências para objetos cujo comportamento muda ao longo do tempo. A detecção de outliers refere-se ao problema de encontrar padrões em dados que são muito diferentes do resto dos dados com base em métricas apropriadas. Esse padrão geralmente contém informações úteis sobre o comportamento anormal do sistema descrito pelos dados. Os algoritmos de distância calculam as distâncias entre os objetos nos dados com interpretação geométrica. Algoritmos baseados em densidade estimam a distribuição de densidade do espaço de entrada e identificam outliers como aqueles que estão em baixa densidade (CHEN et al., 2015).

2.3. Big Data

Big Data constitui-se em uma arquitetura desenhada para trabalhar de maneira econômica na extração do valor de grandes volumes de dados, provenientes de uma variedade de fontes, permitindo alta velocidade na captura, exploração e análise dos dados. Apresenta cinco características principais(SANTANA et al., 2016):

- **Volume:** grandes volumes de dados na ordem de TB (terabytes), PB (petabytes) ou ZB (zettabyte). são gerados, coletados, armazenados e processados;
- **Variedade:** as fontes de dados são heterogêneas e de origens diversas;
- **Velocidade:** a velocidade de geração dos dados ocorre a números cada vez maiores;
- **Veracidade:** garantia de que os dados coletados são confiáveis e consistentes;
- **Valor:** trazer valor financeiro, valor na pesquisa.

A expansão do Big Data e a evolução das tecnologias da Internet das Coisas (IoT) permitem o surgimento de um ambiente onde grandes quantidades de dados coletados por várias fontes permitindo a integração de sensores e dispositivos inteligentes, utilizando serviços altamente conectados em rede. Essa combinação da IoT e Big Data é uma área de pesquisa promissora, a qual trouxe novos e interessantes desafios para alcançar o objetivo das futuras Cidades Inteligente (ABAKER et al., 2016).

2.3.1. Principais Ferramentas de Big Data

Atualmente uma das principais ferramentas de Big Data é o Apache Hadoop, um projeto de alto nível da empresa Apache Software Foundation, que é atualizado frequentemente por contribuintes em uma plataforma em código aberto na linguagem de programação JAVA. Ele foi criado pelo Yahoo em 2006 baseado no Map Reduce, um paradigma da programação introduzido pelo Google para processar e armazenar grandes quantidades de dados, por meio de clusters, onde os mesmos são processados paralelamente e podem ser executados em servidores sem muito esforço. A razão para a escalabilidade desse paradigma é a natureza intrinsecamente distribuída do funcionamento da solução. Uma grande tarefa é dividida em várias microtarefas que são executadas em paralelo em máquinas distintas e então combinadas para chegar à solução da tarefa maior que deu início a todas as outras (RATHORE et al., 2016).

Alguns exemplos de uso do Hadoop são: analisar padrões de usuários para posteriormente trabalhar com sistemas de recomendação, como, por exemplo, a sugestão de novos produtos a serem comprados em sites de e-commerce. A NASA, o Twitter e o Netflix são grandes empresas que utilizam essa ferramenta.

Existem dois componentes principais no Hadoop: Hadoop Distributed File System (HDFS), que é o armazenamento de arquivos, e o MapReduce.

- **HDFS:** Os arquivos grandes são distribuídos entre as máquinas (DataNode) do cluster. Esses arquivos são divididos em blocos de 64MB, em geral, e enviados para as máquinas. Esses blocos são replicados em máquinas diferentes (em geral são 3 máquinas) fato que possibilita a recuperação de dados se necessário. O NameNode é o índice que têm todas as informações dos arquivos, localização e quantidade de blocos. Caso haja alguma falha, o NameNode é avisado e tenta redistribuir novamente os dados. O NameNodeSecundário é um backup para caso ocorra algum problema no NameNode;
- **Framework para MapReduce:** Framework, modelo programático para Hadoop com duas fases, chamadas de Map e Reduce:
 - Primeiro, a função Mapa (Map) divide os dados de entrada (por exemplo, arquivo de texto longo) em partições de dados independentes que constituem pares de valores-chave;
 - Em seguida, a estrutura MapReduce envia todos os pares de valores-chave para o Mapeador que processa cada um deles individualmente, ao longo de várias tarefas de mapeamento paralelo no cluster. Cada partição de dados é atribuída a um nó de computação exclusivo. O Mapeador gera um ou mais pares de valores-chave intermediários. Nesse estágio, a estrutura é cobrada para coletar todos os pares de valores-chave intermediários, para classificá-los e agrupá-los por chave. Portanto, o resultado é uma grande quantidade de chaves com uma lista de todos os valores associados;
 - Em seguida, a função Reduzir (Reduce) é usada para processar os dados de saída intermediários. Para cada chave única, a função Reduzir agrega os valores associados à chave de acordo com um programa predefinido (ou seja, filtrar, resumir, classificar, hashing, obter média ou encontrar o máximo). Depois disso, ele produz um ou mais pares de valores-chave de saída;

- Finalmente, a estrutura MapReduce armazena todos os pares de valores-chave de saída em um arquivo de saída;
- Yarn Gerenciador de Recursos Distribuídos: ele fornece uma melhor escalabilidade, paralelismo aprimorado e gerenciamento avançado de recursos comparado com o MapReduce. Oferece funções do sistema operacional para aplicativos analíticos de Big Data. A arquitetura do Hadoop foi alterada para incorporar o YARN Resource Manager. Em geral, o YARN funciona no topo do HDFS. Essa posição permite a execução paralela de vários aplicativos, permitindo o processamento em lote e o processamento interativo em tempo real. YARN é compatível com Application Programming Interface (API) do MapReduce. Ele garante a contabilidade de tarefas e mantém contadores, tarefas com falha ou lentidão. Assim, a entidade de agendamento de tarefas garante o gerenciamento do ciclo de vida de todos os aplicativos executados em um cluster (RATHORE et al., 2016).

2.3.2. Limpeza dos Dados

As cinco etapas (Limpeza, Agregação, Codificação, Armazenamento e Acesso) não são novas e são conhecidas no caso do gerenciamento de dados tradicional. O principal desafio no Big Data é o gerenciamento da complexidade da natureza dos dados (velocidade, volume e variedade) e processá-los em um ambiente distribuído com um mix de aplicativos. Para obter resultados confiáveis de análise, é essencial verificar a confiabilidade das fontes e a qualidade dos dados antes de utilizar os recursos. No entanto, as fontes de dados podem conter ruídos, erros ou dados incompletos.

2.3.3. Análise e Gerenciamento de Dados

O Big Data apresenta seu potencial quando utilizado para impulsionar a tomada de decisões. Para que isso ocorra, são necessários processos eficientes que transformem rapidamente grandes volumes de dados dinâmicos em insights significativos.

O processo geral de extrair percepções, insights, do Big Data, apresentado na Figura 3, pode ser dividido em cinco etapas que formam os dois subprocessos principais: Gerenciamento de Dados e Análise (GANDOMI; HAIDER, 2015).

- Gerenciamento de Dados envolve processos e tecnologias de suporte para adquirir e armazenar dados, prepará-los e recuperá-los para Análise;
- Análise (Analytics) refere-se às técnicas utilizadas para analisar e adquirir 'extração de insight' de Big Data. Essas técnicas são baseadas em métodos de aprendizado de máquina. Por exemplo, mineração de texto, análise de áudio e vídeo, análise de mídia social e análise preditiva.

Figura 3 - Processo para extrair Insights do Big Data

Fonte: Autora

Em relação à gestão de dados, houve um crescimento significativo de tecnologias capazes de lidar com os Vs de Big Data. O Volume é suportado pelas arquiteturas distribuídas e escalonáveis da maioria destas soluções. A Variedade exigida para sistemas que poderiam eficientemente acomodar formatos estruturados, semi-estruturados e não estruturados. Por fim, a Velocidade exige o desenvolvimento de sistemas quase em tempo real. Nesse cenário, o processamento paralelo massivo (MPP) surgiu como solução. NoSQL e bancos de dados em memória, e o Apache Hadoop são alguns exemplos dessas soluções de Big Data (GANDOMI; HAIDER, 2015).

2.4. Aplicações de Mineração de Dados na Governança das Cidades

No governo das cidades, a Mineração de Dados pode ser usada para detectar as necessidades da população e também para melhorar a qualidade dos serviços públicos, reduzindo os custos e ampliando a participação política com programas mais eficazes, além de promover uma melhora significativa na comunicação entre os cidadãos e o governo. Como exemplo, o sistema de gerenciamento de informações de incidentes da cidade poderá integrar métodos de mineração de dados para fornecer uma avaliação abrangente do impacto de desastres naturais classificando as áreas afetadas, ajudando os governos na preparação de um possível desastre e alocando recursos (CHEN et al., 2015).

Outro exemplo seria a análise de dados criminais, onde os pesquisadores introduzem a técnica de mineração de dados espaciais para descobrir as regras de associação entre os pontos críticos do crime e as localizações. Outros pesquisadores também aproveitam o algoritmo k-means de clustering aprimorado para descobrir padrões de crime e usam técnicas de aprendizado semi-supervisionadas para a descoberta de conhecimento que ajuda a aumentar a precisão da previsão. Além disso, a Mineração de Dados pode ser utilizada para detectar fraudes de identidade criminal, analisando informações de pessoas como nome, endereço, data de nascimento e documentos de identificação para descobrir padrões estruturais previamente desconhecidos de redes criminosas (CHEN et al., 2015).

O Big Data e a Computação na Nuvem são resultados de um sistema computacional poderoso formado por: virtualização, capacidade de armazenamento, redes de banda larga e outros. O cruzamento dessas duas tecnologias envolve o uso de serviço de computação em nuvem e suporte para pesquisa de dados em larga escala, soluções relevantes para as Cidades Inteligentes.

3. REDES SOCIAIS

Conforme Boyd e Ellison (2010), uma rede social é definida como serviços baseados na web, que permitem que os usuários construam um perfil público ou semipúblico em um sistema delimitado, permitindo articular uma lista de outros usuários com quem compartilham conexões e sendo possível conhecer uma lista de conexões feitas por outras pessoas dentro do sistema, visualizar e percorrer sua lista de conexões e aquelas feitas por outras pessoas dentro do sistema. A natureza e a nomenclatura dessas conexões podem variar de site para site.

Atendendo a essas premissas, o primeiro site reconhecido como rede social foi lançado em 1997, cujo o nome é SixDegrees, que permitiu aos usuários criar uma página de perfil, adicionar seus amigos e, em 1998, passou a permitir que os usuários navegassem na listas de seus amigos (BOYD; ELLISON, 2010).

Dentre os anos de 1997 e 2001, o SixDegrees manteve-se ativo e foram registrados 3,5 milhões de usuários em seu auge.

A partir dos anos 2000, ocorreu um grande avanço das Redes Sociais. O Friendster foi um dos precursores desse novo momento. Na sequência, surgiram o Hi5, o LinkedIn e o MySpace, Redes Sociais de nichos específicos, segmentadas aos interesses de seus usuários.

Twitter, Facebook e Pinterest surgiram na sequência, e ainda hoje contemplam a maior parte de usuários de Redes Sociais do planeta. Nessas redes, as pessoas podem compartilhar ideias e sentimentos com seus amigos e seguidores, nos mais diversos formatos, como textos, fotos e vídeos.

A disseminação das Redes Sociais e a conectividade aprimorada possibilitou que as pessoas de diferentes faixas etárias, etnia, status social ou econômico revelassem sobre si mesmas no ambiente on-line. Dados coletados de redes sociais on-line (Online Social Networks OSN) fornecem informações sociais, econômicas e culturais que podem ser utilizadas por governos, formuladores de políticas, autoridades e até mesmo indústrias comerciais para entender melhor as tendências de mercado e padrões de comportamento que podem influenciar a dinâmica individual por meio de dados abertos (MOUSTAKA et al., 2018).

As redes sociais são um fenômeno crescente no consumo atual de mídia na Internet. As redes sociais são usadas não apenas para manter contato com amigos e familiares, mas também para buscar e receber informações sobre produtos/serviços específicos, bem como ativismo social.

As redes sociais, como o Facebook, o Twitter, Instagram etc., permitiram que os indivíduos influenciassem as decisões uns dos outros por meio de uma rede de amizades e assinaturas. A Internet e o mundo físico não são entidades separadas e não relacionadas; eventos no mundo real influenciam as atividades na Internet e vice-versa.

3.1. Cenário das Redes Sociais no Mundo

No início de 2019, o número de usuários de Redes Sociais no mundo alcançou cerca de 3,5 bilhões, com 288 milhões de novos usuários nos últimos 12 meses, elevando a taxa de penetração global para 45% (KEMP, 2019).

No entanto, as Redes Sociais ainda estão longe de serem distribuídas uniformemente pelo mundo, e em especial, em algumas regiões da África, as taxas de penetração ainda são muito baixas.

Verifica-se também que 42% da população mundial acessa as Redes Sociais diretamente por dispositivos móveis. Um usuário de Redes Sociais gasta, em média, 2 horas e 16 minutos por dia em plataformas sociais, o que equivale a aproximadamente um terço do tempo total de navegação na Internet e um sétimo do tempo que uma pessoa fica acordada em um dia (KEMP, 2019).

O Facebook (com 2,2 bilhões de usuários) e YouTube (com 1,9 bilhão de usuários) são as maiores plataformas de Redes Sociais da atualidade. O Twitter, utilizado como fonte de dados para este trabalho, tem atualmente 326 milhões de usuário, destacando-se entre as 10 maiores Redes Sociais do mundo (KEMP, 2019).

3.2. Redes Sociais e Cidades Inteligentes

As tendências demográficas, ambientais, econômicas e tecnológicas, combinadas com a sustentabilidade urbana, levaram à concepção e desenvolvimento de Cidades Inteligentes capazes de enfrentar desafios recentes e futuros, aproveitando as tecnologias de informação e comunicação (TIC).

A aquisição de conhecimento urbano, que vem da exploração de dados urbanos, ajuda as cidades a identificar as fraquezas e oportunidades locais e determina a tomada de decisões e a implantação de serviços inteligentes, enquanto elas dizem respeito aos principais pré-requisitos para transformar uma cidade comum em uma Cidade Inteligente.

Uma grande variedade de dispositivos fixos ou portáteis (por exemplo, sensores, câmeras, medidores, atuadores e RFID, etc.), a chamada Internet das Coisas (IoT) e aplicativos (por exemplo, OnLine Social Networks, plataformas web, aplicações móveis, etc.) foram desenvolvidos e utilizados para capturar diferentes aspectos da vida nas cidades (MOUSTAKA et al., 2018).

Os smartphones levaram à rápida expansão e aumento do uso das Redes Sociais, nas quais uma enorme quantidade de fluxos de dados urbanos heterogêneos, gerados a partir das fontes de dados mencionadas anteriormente, resultam na transição do “urbanismo informado por dados” para o “urbanismo orientado por dados”, conforme apropriadamente apontado por (KITCHIN apud MOUSTAKA et al., 2018).

Algumas pesquisas já vêm sendo realizadas utilizando a Análise de Sentimentos a partir dos dados gerados em Redes Sociais. Em 2012, por exemplo, durante a Olimpíada de Londres, na Inglaterra, a London Eye, roda-gigante e ponto turístico da cidade, recebeu uma iluminação que variava as cores de acordo com os sentimentos da população referentes aos jogos olímpicos, captados via Twitter. O "acompanhamento do sentimento social" em tempo real dividiu os tweets em conversas positivas e negativas e os filtrou por meio de um programa que sistematicamente os convertia em um show de luzes (BBC, 2012).

Outro uso recente da Análise de Sentimentos em Redes Sociais aconteceu durante as candidaturas de Donald Trump e Hillary Clinton para as eleições dos Estados Unidos, em 2016. Durante o período de candidatura, milhões de tweets publicados pela população, mencionando um dos candidatos foram analisados pelo sistema MonkeyLearn. Esse sistema

era capaz de avaliar o sentimento de cada Tweet como positivo, neutro ou negativo (STECANELLA, 2016).

4. CIDADES INTELIGENTES (SMART CITIES)

Devido à tendência mundial de crescimento da população das cidades, citado anteriormente, novos desafios surgirão para os seus gestores, como problemas de tráfego, segurança, educação, saúde, consumo de água e energia, entre outros. Com isso, as Tecnologias da Informação e Comunicação (TIC) podem se tornar ferramentas importantes no auxílio ao monitoramento, controle e tomada de decisões diante de tais situações (REDE BRASILEIRA, 2016).

Todo esse contexto favorece o desenvolvimento de Cidades Inteligentes, com o objetivo de otimizar a utilização de recursos e infraestrutura para uma maior sustentabilidade, qualidade de vida da população, com coesão social, conectividade e inovação(BERRONE et al., 2018).

Por definição, é esperado que essas cidades utilizem tecnologias avançadas de interação social e uma infraestrutura tecnológica resiliente, interoperável e transparente para a geração e o gerenciamento de dados de forma aberta e acessível e em constante melhoria e evolução, o que permite melhorar, aumentar e automatizar as funções da cidade de forma eficiente, integrada, sustentável e relevante para a população (PEREIRA et al., 2017).

Seguem alguns cenários de computação urbana a serem monitorados para o planejamento das Cidades Inteligentes (TOBERGTE; CURTIS, 2013):

- Meio ambiente: através de redes de sensores é possível fazer o monitoramento exato de condições ambientais, como níveis de poluição, controles de animais selvagens, escoamento de água (chuva e esgoto) e outros;
- Proteção: monitoramento estrutural de edifícios, pontes e barragens, bem como sistemas de alerta avançado em situações de emergência, como vazamento de uma usina nuclear e outros;
- Transporte: através de sensores incorporados em vias e postes, o tráfego pode ser monitorado em tempo real com a finalidade de reduzir o tempo de viagem e a economia de combustível. É possível também monitorar e identificar a disponibilidade em tempo real de vagas de estacionamento através de dispositivos inteligentes;
- Serviços e Utilidades: uma rede de serviços públicos inteligente capacitará os usuários finais a serem mais conscientes de seus consumos de energia e abastecimento, permitindo que as empresas forneçam apenas a quantidade de energia ou água

necessária. Por exemplo, o sistema instalado pela IBM em Malta que integra sistemas de água e energia e é capaz de identificar vazamentos de água e perdas de eletricidade na rede. Ao todo, cerca de 250.000 medidores interativos monitoram o uso em tempo real, estabelecendo taxas variáveis e recompensando os clientes que consomem menos recursos;

- Construções Inteligentes: os Edifícios Inteligentes utilizam dispositivos de monitoramento que rastreiam o uso e capacitam os usuários e provedores de serviços para melhor controlar e reduzir as demandas de eletricidade e abastecimento. Por exemplo, o controle de temperatura do ar-condicionado pode aprender os padrões de uso dos cidadãos e adaptar aos seus comportamentos e preferências ao longo do tempo.

4.1. Internet das Coisas (IoT), Big Data e as Cidades Inteligentes

A infraestrutura de uma Cidade Inteligente está relacionada com a disponibilidade de grandes dados existentes em diversos formatos (Big Data) e com a interconexão de todos os componentes da cidade (IoT). Sensores de poluição, sistemas de tráfego, conteúdos postados em mídias sociais e dados de smartphones são alguns exemplos de componentes que geram uma quantidade crescente de dados de todos os tipos, públicos ou privados, estruturados ou não-estruturados, transmitidos ou estáticos. Com o objetivo de fortalecer e aprimorar a governança de uma Cidade Inteligente, esses dados devem ser coletados, tratados, analisados e visualizados (AHMED et al., 2016).

O avanço tecnológico da Internet e o grande crescimento no número de dispositivos inteligentes capazes de monitorar, processar, comunicar e atuar promovem o contexto ideal para o surgimento da Internet das Coisas (Internet of Things - IoT), onde todos esses dispositivos ou objetos conectados na Internet podem promover a comunicação entre si, entre os usuários e entre usuários e dispositivos.

A cada dia, o número de dispositivos ou objetos (coisas) conectados à Internet aumenta, sendo previsto cerca de 20 bilhões de unidades conectadas em 2020. Uma nova geração de aplicativos e serviços no ambiente da Internet das Coisas (IoT) está surgindo, alimentada por dados e recursos dos mundos físico e virtual (GIANG et al., 2016).

A Internet das Coisas (IoT) é uma das tecnologias mais promissoras para enfrentar o desafio da expansão da urbanização nas cidades modernas e exige soluções inteligentes para abordar questões críticas como mobilidade, saúde, energia e infraestrutura civil, entre outras (ALAVI et al., 2018).

Um dos desafios das Cidades Inteligentes para obter um planejamento urbano eficiente é a integração dos milhares de dispositivos IoT interconectados, conforme a, Figura 4, que produzem um enorme volume de dados, denominados Big Data cujos dados precisam ser processados (RATHORE; AHMAD; PAUL, 2015).

Figura 4 - Implantação de Sensores e Geração de Dados do Sistema Inteligente

Fonte: Autora

A combinação da IoT com o Big Data é uma área de pesquisa pouco explorada e que permite o surgimento de novos desafios com o intuito de alcançar os objetivos das futuras Cidades Inteligentes (RATHORE; AHMAD; PAUL, 2015).

A análise e a utilização efetiva de Big Data são fatores-chave para o sucesso em muitos domínios de negócios e de serviços. A aplicação de Big Data em uma cidade apresenta muitos

benefícios e desafios, incluindo a disponibilidade de grandes instalações computacionais e de armazenamento para processar fluxos de dados produzidos em um ambiente de Cidade Inteligente (ABAKER et al., 2016).

Há muitos aplicativos de análise de Big Data em diferentes domínios de Cidades Inteligentes, como planejamento de investimentos por região, controle de tráfego e transporte, análise de segurança, distribuição energética e monitoramento ambiental. A heterogeneidade das fontes de dados detectados nos diversos segmentos da sociedade envolve fatores de design consideráveis, como integração de dados, escalabilidade do sistema, privacidade e segurança. Além disso, a natureza dinâmica da vida na Cidade Inteligente exige atenção considerável para a análise do fluxo de dados, permitindo serviços online e em tempo real.

A análise de “dados históricos” possibilita a extração de insights e correlações ocultas nos dados (Big Data), sendo um requisito essencial para o planejamento (curto e longo prazo) e para fins de tomada de decisão. No entanto, para extrair insights valiosos para o desenvolvimento de serviços de informações inteligentes em nível de cidade, os conjuntos de dados gerados de vários domínios da cidade precisam ser integrados e analisados. Esse processo geralmente é chamado de análise de Big Data ou Cadeia de Valores de Big Data (OSMAN, 2019).

4.2. Smart Human Cities (SHC)

O conceito que vem sendo chamado de Cidades Inteligentes (Smart Cities) é uma tendência mundial relativa ao emprego de práticas de sustentabilidade e ao uso de soluções intensivas de TICs (Tecnologias da Informação e Comunicação) como instrumentos para tornar as cidades mais inteligentes, otimizando os recursos para uma melhor qualidade de vida da população.

Além da visão tradicional ligada ao uso de TICs para as Cidades Inteligentes, buscam-se incluir aspectos ligados ao cidadão, já que a cidade é composta de pessoas, por pessoas e para pessoas. As discussões visam questionar como os cidadãos têm se engajado, o que eles necessitam no que concerne à informação e como é possível humanizar o uso de TICs em busca da melhoria de qualidade de vida em centros urbanos. Assim, o conceito evolui de Cidades Inteligentes (Smart Cities - SC) para Cidades Inteligentes e Humanas (Smart Human Cities - SHC) (REDE BRASILEIRA, 2016).

O conceito Smart Human Cities (SHC), conforme Figura 5, é aquele que sustenta sua própria evolução contínua, tendo como objetivos o bem-estar, a qualidade de vida e a coparticipação do cidadão e das comunidades locais, apoiando seu desenvolvimento em ações, projetos e políticas públicas que promovam (de forma igualitária) colaboração entre comunidade, autoridades públicas e sociedade civil para mediação e resolução de conflitos, bem como promoção da criatividade local.

Figura 5 – Elementos-chave das Cidades Inteligentes e Humanas

Fonte: Autora adaptada de (PEREIRA et al., 2017)

Neste trabalho, utiliza-se o conceito de Cidades Inteligentes e Humanas (SHC). A participação dos cidadãos nesse conceito possibilita que eles sejam tratados como fontes de informações similares aos dispositivos e sistemas que captam dados na Cidade Inteligente. Assim, os cidadãos passam a ser considerados parte do grupo de sensores, com o diferencial de não haver custo adicional para compra e implantação, sendo que suas informações chegam em tempo real, possuem mobilidade e dinamismo.

Um dos meios ligados aos Fatores Humanos e capaz de contribuir com o desenvolvimento desse processo é o conteúdo compartilhado por usuários em Redes Sociais, fontes de dados

ricas e que podem ser tratadas em Análise de Sentimentos. Além dos dados captados em forma escrita, alguns trabalhos recentes tentaram prever o sentimento visual dos usuários usando características das imagens por eles publicadas (AHMED et al., 2016).

O desenvolvimento de dispositivos móveis, especialmente smartphones com grande poder computacional, onde diferentes tipos de sensores são adicionados, como acelerômetro, giroscópio, bússola, microfone, câmera, GPS e rede sem fio capacitam os seres humanos a sentir o ambiente através dessas interfaces de sensores inteligentes. Essa tecnologia dá suporte ao Mobile Crowd Sensing (MCS) que é definido como um novo paradigma de sensoriamento, que capacita cidadãos comuns a contribuir com dados detectados oportunamente ou gerados pelos próprios usuários a partir de seus dispositivos móveis e agrega e funde os dados na nuvem para extração de inteligência e prestação de serviços centrada no ser humano (XU et al., 2018).

4.3. Estudos de Casos das Cidades Inteligentes

Nesta seção são descritos alguns exemplos de Cidades Inteligentes da atualidade:

- A Cidade de Helsinki na Finlândia, teve seu desenvolvimento como uma Cidade Inteligente inspirado pela existência de dados públicos abertos. O Projeto Infoshare da Região de Helsinki é composto por uma das plataformas pioneiras de dados urbanos abertos, onde as informações podem ser acessadas por indivíduos da população, setor acadêmico, governo, empresas e instituições de pesquisa. Em 2013, foram disponibilizadas mais de 1030 bases de dados que contemplam informações referentes à cidade, como transporte, economia, condições, emprego e bem-estar. A plataforma foi recentemente premiada com o Prêmio Europeu de Inovação na Administração Pública na categoria de capacitar os cidadãos. O relatório do júri sugere que a abertura de informação sobre a tomada de decisões através de um sistema eletrônico de gestão de casos proporciona aos cidadãos uma grande oportunidade de estarem significativamente envolvidos na tomada das decisões públicas (ABAKER et al., 2016);

- Copenhague se comprometeu atingir a emissão neutra de carbono até 2025. A população da cidade utiliza a bicicleta regularmente como meio de transporte, cerca de 40% dos munícipes, e quando não há trânsito nas ruas, os semáforos se desligam automaticamente para economizar energia. Além disso, medidores de qualidade do ar foram integrados aos postes de iluminação pública, e a energia solar é utilizada para iluminar espaços públicos. Copenhague ainda investiu no desenvolvimento de ferramentas para facilitar o uso da cidade, como aplicativos para encontrar vagas de estacionamento disponíveis nas ruas e gerenciadores de consumo de energia residencial (ABAKER et al., 2016);
- Em 2018, Nova York foi eleita a cidade mais inteligente do mundo de acordo com a quinta edição do IESE Cities in Motion Index (CIMI), escola de negócios IESE da Universidade de Navarra, na Espanha. A cidade lançou uma plataforma interativa em colaboração com a Cisco que converteu sistemas telefônicos públicos antigos para fornecer acesso à internet para todos os residentes. A plataforma oferece alertas de segurança, informações sobre eventos locais, notícias de vizinhança e listas de entretenimento (BERRONE et al., 2018). Além disso em 2017, foi realizada a instalação de sensores de presença em 90 escolas da cidade com o intuito de economizar 17 milhões de kWh, reduzindo as emissões de gases do efeito estufa e economizando mais de US\$ 2 milhões por ano. O Departamento de Transportes passou a gerenciar o tráfego da cidade a partir de dados em tempo real com 300 sensores e câmeras capazes de fornecer estatísticas e modificar os padrões dos semáforos em uma ampla região, resultando em uma melhora de 10% nos tempos de viagem desde que o programa foi implementado em 2010.
- São Paulo (SP) é a cidade mais inteligente do Brasil, de acordo com ranking nacional Cidades em Movimento 2018, divulgado com exclusividade pela escola de negócios IESE da Universidade de Navarra, na Espanha. Apesar de estar na liderança, a capital paulista perde o pódio em todos os indicadores de forma isolada: economia, capital humano, aspectos sociais, meio ambiente, governança corporativa, planejamento urbano, acesso a serviços internacionais, tecnologia e mobilidade e transporte. A

cidade de São Paulo, além de ser a mais populosa do Brasil, é a que possui o maior número de veículos e meios de transporte. É também uma das cidades que possuem uma empresa específica para cuidar da mobilidade dos cidadãos, a CET (Companhia de Engenharia de Tráfego), prestadora de serviços oficial da prefeitura de São Paulo. Atualmente, a cidade de São Paulo é uma das que mais sofrem com problemas de mobilidade urbana. De acordo com levantamento, o paulistano gasta, em média, três horas diariamente no trânsito de São Paulo levando em conta todos os deslocamentos pela cidade. E a pesquisa mostra que a maioria da população paulistana utiliza o transporte coletivo: 47% usam ônibus; 22%, carro; 13%; metrô, 8% andam a pé; 4% utilizam trem; 2%, transporte particular como Uber; 1%, motocicletas; e 1%, bicicleta (BOCCHINI, 2019).

O tema Transporte Urbano é utilizado para a realização do estudo de caso apresentado neste trabalho.

5. ANÁLISE DE SENTIMENTOS

Opiniões e sentimentos desempenham um papel importante em quase todas as ações humanas, influenciando o modo como pensam, o que fazem e como agem (HAJMOHAMMADI; OTHMAN, 2012).

A Análise de Sentimentos, também conhecida como Mineração de Opinião, é o campo de estudo que analisa as opiniões, sentimentos, avaliações, atitudes e emoções das pessoas em relação a entidades como produtos, serviços, organizações, indivíduos, questões, eventos e outros (LIU, 2012).

A Análise de Sentimentos é uma das áreas de pesquisas em Processamento de Linguagem Natural (PLN) que mais tem sido abordada desde o ano 2000. Dentre suas particularidades, ela tem se destacado por sua influência em diversas áreas, como ciências de gestão, ciência política, economia e ciências sociais, áreas diretamente ligadas às opiniões das pessoas (LIU, 2012).

O Processamento de Linguagem Natural (PNL) é uma das áreas da ciência da computação que trabalha com idiomas utilizando técnicas para processar a linguagem humana. A PNL incorpora muitos subcampos e tarefas, alguns dos quais são resumos automáticos, análise de discurso, tradução automática, extração de relacionamento e respostas a perguntas. O estudo nessa área tem uma repercussão positiva para os desenvolvimentos de outras áreas em diferentes campos (HOPFGARTNER, 2015).

O objetivo da Análise de Sentimentos é definir ferramentas automáticas capazes de extrair informações subjetivas de textos em linguagem, tais como sentimentos de opinião, de modo a criar conhecimento estruturado e acionável para ser usado em tomadas de decisões (POZZI; FERSINI; LIU, 2017).

Este processo envolve técnicas de diversas áreas como Mineração de Textos, Aprendizado de Máquina e Processamento de Língua Natural. A Mineração de Opiniões é uma área de pesquisa relativamente nova, mas vem sendo amplamente estudada e apresenta diversos desafios (LIU, 2012).

Uma das primeiras pesquisas de Análise de Sentimentos foi realizada por Pang, Lee e Vaihyanathan em 2002 para análise de sentimentos relacionada a filmes, onde foram utilizadas técnicas de aprendizado de máquina (PANG; LEE; VAITHYANATHAN, 2002).

As opiniões, na maioria das vezes, são o centro de quase todas as atividades humanas e, conseqüentemente, influenciadoras de comportamentos. Quando precisamos tomar uma decisão, muitas vezes buscamos saber a opinião de outras pessoas. Na atualidade, as empresas e organizações querem saber as opiniões dos consumidores referentes aos seus produtos e serviços. Os consumidores também querem saber as opiniões dos usuários de um determinado produto ou serviço antes de adquiri-lo.

A Internet é um ambiente que possibilita a expressão de sentimentos e opiniões sobre os mais variados assuntos, por meio de posts em redes sociais, blogs, fóruns online, avaliações de produtos em sites, etc. Os usuários são capazes de fornecer informações e influenciar uns aos outros. Assim, um grande volume de dados rico em opiniões e sentimentos está sendo gerado na Web diariamente e pode ser analisado (LIU, 2012).

As opiniões expressadas em redes sociais, por exemplo, podem ser extraídas e analisadas de forma eficiente para serem utilizadas nas tomadas de decisões em tempo real. Elas podem auxiliar no entendimento de diversos fenômenos sociais complexos e também prevêê-los. Nesse intuito, são realizadas pesquisas utilizando técnicas de mineração de dados, processamento de linguagem natural e aprendizagem de máquina (RIBEIRO, 2013).

Num futuro próximo, é esperada a chegada da Web 5.0, Web sensorial e emotiva, onde é previsto que todas as informações disponíveis na rede sejam captadas e analisadas para a criação de sistemas que facilitem a comunicação dos dados entre humanos, máquinas e objetos em ambientes ubíquos. Visando auxiliar esse processo, estão sendo propostos sistemas que tratam as opiniões de forma automática por meio dos conceitos da área de Análise de Sentimentos (AS), também conhecida por Mineração de Opinião (ZANELLA et al., 2014).

5.1. Termos e Conceitos Utilizados em Análise de Sentimentos

Nesta seção são definidos os principais termos e conceitos utilizados na técnica de análise de sentimentos.

5.1.1. Definições

A Análise de sentimentos ou Mineração de Opiniões opera sobre porções de texto que apresentem uma opinião referente a uma determinada entidade. O objetivo é identificar os sentimentos encontrados em relação a essa entidade. A entidade pode ser um produto, uma pessoa, um evento, um serviço, uma empresa, etc.

Toda opinião é composta de pelo menos dois elementos chave: um alvo e um sentimento sobre este alvo. Onde um alvo pode ser uma entidade ou aspecto da entidade e o sentimento representa uma atitude, opinião ou emoção que o autor da opinião tem a respeito do alvo. A polaridade de um sentimento corresponde a um ponto em alguma escala que representa a avaliação positiva, neutra ou negativa do significado deste sentimento (BECKER; TUMITAN, 2013).

Dado um texto, a tarefa inicial na mineração de opiniões consiste em determinar se o mesmo é subjetivo, ou seja, expressa um sentimento. Se for considerado subjetivo, uma opinião pode ser representada como uma quintupla $O = (e, a, s, h, t)$, onde (LIU, 2010) (LIU, 2012):

- e: é o nome de uma entidade;
- a: é um aspecto da entidade e (opcional);
- s: é a polaridade do sentimento sobre aspecto a que tem como alvo a entidade e;
- h: é o detentor do sentimento (i.e. quem expressou o sentimento), também chamado de fonte de opinião;
- t: é o instante no qual a opinião foi expressa por h.

O conceito de aspecto, também denominado característica ou propriedade, permite que uma entidade seja vista através de diferentes perspectivas ou atributos. Como exemplo, considere o sentimento expresso no comentário realizado por Fernando.

Fernando - 6 de julho de 2018: “Adorei Conhecer São Paulo. Excelentes restaurantes e boas opções de lazer. Pena que tem muito trânsito.”

Esse comentário será representado por quatro quintuplas, sendo que uma refere-se à cidade de São Paulo e as outras aos aspectos da cidade.

Os aspectos da Cidade de São Paulo são: os restaurantes, o lazer e o trânsito.

- (São Paulo, geral, positivo, Fernando, 06/07/2019);
- (São Paulo, restaurantes, positivo, Fernando, 06/07/2019);
- (São Paulo, lazer, positivo, Fernando, 06/07/2019);
- (São Paulo, trânsito, negativo, Fernando, 06/07/2019).

Os termos sentimento e opinião frequentemente são usados como sinônimo neste contexto.

5.1.2. Polaridade

Representa o grau de positividade (satisfação), neutralidade ou negatividade (insatisfação) de um sentimento expresso por um usuário perante determinado tema, produto ou serviço. Alguns métodos classificam a polaridade de um sentimento em classes discretas do tipo binário (positivo ou negativo) ou tipo ternário (positivo, negativo ou neutro). Como exemplo, veja a Tabela 1:

Tabela 1 – Exemplos de frases e suas respectivas polaridades

INFORMAÇÃO (FRASE)	POLARIDADE
As fotos são maravilhosas!	Positiva
O sinal do WiFi é péssimo!	Negativa
Eu não sei se eu gostei do filme ou não.	Neutra

Fonte: Autora

A polaridade de um sentimento também pode ser representada por um intervalo numérico que representa a intensidade deste sentimento, normalmente é representado por um ponto flutuante entre intervalos de -1 e +1, -5 e +5 ou até $-\infty$ e $+\infty$.(NIELSEN, 2011).

5.1.3. Categorização das Sentenças: Subjetividade x Objetividade

As sentenças podem ser divididas em objetivas e subjetivas. As objetivas apresentam normalmente um fato ou uma informação. Se uma sentença é classificada como objetiva, nenhuma outra tarefa fundamental é necessária, enquanto se a sentença for classificada como subjetiva, sua polaridade (positiva, negativa ou neutra) precisa ser estimada, Figura 6 (POZZI; FERSINI; LIU, 2017).

As sentenças subjetivas expressam sentimentos pessoais e opiniões. Cabe ressaltar que textos informais (ex.: coletados de redes sociais) tendem a ser mais subjetivos que textos formais (ex.: coletados de notícias) (RIBEIRO, 2013).

Figura 6 - Categorias de Sentenças

Fonte: Autora

Um exemplo de sentença objetiva é “O iPhone é um smartphone”, enquanto um exemplo de sentença subjetiva é “O iPhone é incrível”. A classificação de polaridade é a tarefa que distingue sentenças que expressam polaridades positivas, negativas ou neutras. Uma sentença subjetiva pode não expressar qualquer sentimento positivo, negativo ou neutro. Esta sentença é classificada como neutra “Eu acho que ele chegou” (POZZI; FERSINI; LIU, 2017).

5.1.4. Tipos de Opinião ou Análise de Sentimentos

As opiniões podem ser divididas em regular e comparativa. Uma opinião regular traz informações de maneira direta ou indireta sobre determinado assunto (LIU, 2012).

- **Opinião regular:** uma opinião regular é frequentemente referida na literatura como um padrão de opinião e tem dois subtipos principais:

- **Opinião direta:** uma opinião direta refere-se a uma opinião expressa diretamente a uma entidade ou aspecto (por exemplo, “O **brilho da tela** do iPhone é incrível”);
- **Opinião indireta:** uma opinião indireta é uma opinião que é expressa indiretamente a uma entidade ou aspecto com base em seus efeitos sobre algumas outras entidades. Por exemplo, a frase “Depois que mudei para o iPhone, perdi todos os meus dados!” Descreve um efeito indesejável da opção “ meus dados ”, que indiretamente dá um sentimento negativo ao iPhone.

- **Opinião comparativa:** uma opinião comparativa expressa uma relação de semelhanças ou diferenças entre duas ou mais entidades e/ou uma preferência do detentor de opinião com base em alguns aspectos compartilhados das entidades. Por exemplo, as frases “iOS tem melhor desempenho que o Android” e “iOS é o sistema operacional com melhor desempenho” expressam duas opiniões comparativas. Uma opinião comparativa é geralmente expressa com o uso da forma comparativa ou superlativa de um adjetivo ou advérbio. O objetivo desta Análise de Sentimento é identificar as sentenças que contêm as opiniões para serem comparadas (POZZI; FERSINI; LIU, 2017).

5.1.5. Semântica

A semântica da linguagem usada principalmente nas redes sociais é fundamental para analisar, com precisão, as expressões do usuário. O contexto de uma expressão textual é um elemento crucial que deve ser levado em conta para lidar adequadamente com o sentimento subjacente. Uma sentença “tomada como está” pode parecer negativa ou positiva, mas se for adequadamente analisada do ponto de vista semântico, pode ser completamente diferente. Por exemplo, as frases “Eu assisti ao filme de terror mais fantástico. Foi como um verdadeiro pesadelo! MEEEDDDOOO” pode ser inicialmente interpretado como negativo, mas levando em consideração o contexto, onde esses tipos de opiniões são expressos numa comunidade de apreciadores de filmes de terror, deve ser interpretado como positivo. Lexicos, corpora e ontologias precisam ser adequadamente construídos e usados para que possamos ter uma compreensão profunda da semântica da linguagem natural nas redes sociais online (POZZI; FERSINI; LIU, 2017).

5.1.6. Lidando com Figuras de Linguagem

As figuras mais problemáticas do discurso no Processamento da Linguagem Natural (PLN) são a ironia e o sarcasmo. Enquanto a ironia é frequentemente usada para enfatizar ocorrências que se desviam do esperado, como reviravoltas do destino, o sarcasmo é comumente usado para transmitir críticas implícitas a uma determinada vítima como alvo. A característica inerente aos atos de fala sarcásticos e irônicos é que às vezes eles são difíceis de serem reconhecidos, seja por humanos e/ou por máquinas. A dificuldade no reconhecimento do sarcasmo e da ironia causa mal-entendidos na comunicação cotidiana e coloca problemas a muitos sistemas de processamento de linguagem natural (POZZI; FERSINI; LIU, 2017).

No contexto da análise do sentimento em geral o sarcasmo e a ironia são considerados como sinônimos. Muitas vezes quando uma sentença sarcástica/irônica é detectada como positiva, provavelmente significa negativa e vice-versa. Não será foco desta pesquisa o tratamento dessas figuras de linguagem.

5.2. Níveis de Análise de Sentimentos

Como já mencionado anteriormente o objetivo da Análise de Sentimentos é “definir ferramentas automáticas capazes de extrair informações subjetivas de textos em linguagem natural”. A primeira escolha a ser definida em uma Análise de Sentimentos é definir o que é o texto, ou seja, o objeto de análise. Em geral, a Análise de Sentimentos pode ser investigada principalmente em quatro níveis, conforme Figura 7:

Figura 7 – Níveis de Análise de Sentimentos

Fonte: Autora

5.2.1. Nível do Documento (Mensagem)

A tarefa neste nível é classificar a opinião de um documento inteiro expressa em um sentimento positivo ou negativo (PANG et. al; 2002).

Por exemplo, em um documento ou Mensagem, dada a revisão de um produto, o sistema determina se a mensagem expressa uma opinião geral positiva, negativa ou neutra sobre o produto. A suposição é que a mensagem inteira expresse apenas uma opinião sobre uma única entidade (por exemplo, um único produto). Este tipo de classificação não é recomendado para a análise de documentos que avaliam e comparam múltiplas entidades.

Este é o nível adotado para análise neste trabalho de pesquisa.

5.2.2. Nível de Sentença

Em um único documento é possível identificar múltiplas opiniões ou mesmo entidades. O texto é dividido em frases ou sentenças que possam conter opiniões individuais. Cada sentença expressa uma opinião positiva, negativa ou neutra, onde neutra não significa opinião (LIU, 2012).

O objetivo é determinar a polaridade de cada frase contida em uma mensagem de texto. A suposição é que cada sentença, em uma mensagem dada, denota uma única opinião em uma única entidade.

5.2.3. Nível de Aspecto

Uma sentença pode ser julgada por várias entidades e pode conter múltiplos sentimentos associados a ela. Por exemplo, a sentença “Esse hotel, apesar de possuir um ótimo quarto, tem um atendimento péssimo!” possui duas diferentes polaridades associadas a “quarto” e “atendimento” para o mesmo hotel. Enquanto “quarto” é considerado positivo, “atendimento” pode ser analisado de forma negativa. Esta necessidade de avaliar a opinião para cada entidade é comum em reviews de produtos ou em fóruns de discussões.

O nível de Aspecto executa uma análise mais detalhada do que os níveis de mensagem e frase. Baseia-se na ideia de que uma opinião consiste em um sentimento e um alvo (de opinião). Por exemplo, a frase “O iPhone é muito bom, mas eles ainda precisam trabalhar com a bateria e questões de segurança” avalia três aspectos: iPhone (positivo), duração da bateria (negativo) e segurança (negativo) (POZZI; FERSINI; LIU, 2017).

5.2.4. Nível de Palavra ou Dicionário

Em alguns estudos, também pode ser encontrado um nível ainda mais aprofundado nos significados das palavras e termos, chamado nível de Palavra ou Dicionário, onde o foco é a otimização dos dicionários léxicos de sentimentos, com atualização de palavras do cotidiano, gírias e acrônimos das redes sociais. Além disso, essas palavras recebem polaridades e intensidades de sentimentos (NIELSEN, 2011).

5.3. Métodos de Classificação para Análise de Sentimentos

Existem duas abordagens principais para a classificação de sentimentos uma baseada em Léxico e a outra em Aprendizado de Máquina.

Uma abordagem baseada em Léxico (um conjunto de recursos linguísticos com semântica bem definida) transforma os dados em palavras individuais que são verificadas com um léxico de sentimento contendo um valor de polaridade para cada palavra. A soma das polaridades é passada para um algoritmo que determina a polaridade geral da sentença (SZERSZEN; PALSSON, 2016).

Uma abordagem de Aprendizado de Máquina utiliza modelos de aprendizado supervisionado e um conjunto de treinamento rotulado para adaptar um classificador ao domínio de dados do conjunto de treinamento. O classificador treinado pode então prever o resultado do problema e a taxa de sucesso da predição dependerá do quão bem o problema está contido no mesmo domínio (SZERSZEN; PALSSON, 2016).

5.3.1. Classificadores Baseados em Léxico

Um léxico só precisa de um documento (uma espécie de dicionário) contendo palavras com sua polaridade e sua orientação semântica, e não precisa ser treinado ou processado antes do uso (TABOADA et al., 2011).

Uma das técnicas de Análise de Sentimentos baseada em léxico emprega a correspondência de dicionário em um léxico predefinido de sentimentos para derivar a polaridade do sentimento. Estes métodos costumam usar um conjunto de regras geralmente definido pelo homem para lidar com a negação da polaridade, conforme Figura 11 (RIBEIRO, 2013).

Os dicionários contendo os léxicos, ao invés de possuírem como conteúdo o significado de cada palavra, possuem em seu lugar um significado quantitativo, que pode ser um número entre -1 a 1, onde -1 é o valor sentimental mais negativo e 1 o valor mais positivo, ou também um valor qualitativo positivo ou negativo, feliz ou triste. Abordagens léxicas assumem que palavras individuais possuem o que é chamado de polaridade prévia, uma orientação semântica independente de contexto e que pode ser expressada com um valor numérico ou classe (TABOADA et al., 2011).

A Figura 8 generaliza o funcionamento de um método de análise de sentimentos léxico. O processo de classificação inicia quando o método recebe uma sentença de entrada, em seguida é realizado um processamento da linguagem natural, assim como uma pesquisa no léxico dos termos que formam esta mensagem. Ao final do processo o método é capaz de inferir qual é a polaridade ou sentimento implícito na sentença de entrada.

Figura 8 – Léxico de Sentimentos

Fonte: Autora adaptada de (RIBEIRO, 2013)

O documento de polaridade é uma lista de palavras e sua respectiva orientação semântica (TABOADA et al 2011). O resultado da classificação de sentimento é apresentado como pontuação binária positiva ou negativa, uma pontuação neutra às vezes também incluída. .

Os dicionários de sentimento usados para a Análise de Sentimentos baseada em léxicos podem ser criados manualmente ou automaticamente usando palavras de semente para expandir a lista de palavras. Por exemplo, o dicionário WordNet lista sinônimos e antônimos para cada palavra. Assim, uma técnica simples nesta abordagem é usar algumas palavras de sentimento para bootstrap com base no sinônimo e estrutura de antônimos de um dicionário. Especificamente, este método funciona da seguinte maneira: um pequeno conjunto de sentimentos, sementes, com orientações positivas ou negativas conhecidas é coletado manualmente.

O algoritmo, em seguida, aumenta este conjunto, pesquisando no WordNet ou outro dicionário online para seus sinônimos e antônimos. As palavras recém-encontradas são

adicionadas à lista de sementes. A próxima iteração começa. O processo iterativo termina quando não mais palavras novas podem ser encontradas. Depois que o processo for concluído, uma etapa de inspeção manual foi usada para limpar a lista.

Os classificadores baseados em léxico são mais simples de criar e implementar do que os baseados em Aprendizado de Máquina, não há a necessidade de um conjunto de treinamento pré-definido. Embora eles não tenham um desempenho tão bom quanto os classificadores baseados em aprendizado, ainda há benefícios em usar um classificador baseado em léxico. Isso deve-se ao fato de que os mesmos não requisitam nenhum conhecimento específico do domínio, diferente dos classificadores baseados em Aprendizado de Máquina que quando usados em um domínio distinto daquele para o quais foram treinados terão uma perda de desempenho, enquanto uma abordagem baseada em léxico não é afetada (SZERSZEN; PALSSON, 2016).

5.3.2. Classificadores baseados em Aprendizagem de Máquina Supervisionada

A Análise de Sentimentos baseada na aprendizagem de máquina utiliza métodos de aprendizagem automática para classificar frases ou documentos em duas (negativas e positivas) ou três (negativas, positivas e neutras) polaridades.

Esta abordagem exige uma etapa de treinamento de um modelo com amostras previamente classificadas. O procedimento para realizar a aprendizagem de máquina compreende quatro etapas principais: obtenção de dados rotulados para treinamento e teste, definição das features ou características padrões que permitirão a distinção entre os dados; treinamento de um modelo computacional com um algoritmo de aprendizagem e por último a aplicação do modelo (RIBEIRO, 2013).

Pang, Lee e Vaithyanathan (2002) investigaram três métodos de aprendizagem automática para produzir classificadores automatizados para gerar rótulos de classe para revisões de filmes. Eles os testaram em Naive Bayes, Maximum Entropy e Support Vector Machine (SVM), e avaliaram a contribuição de diferentes características, incluindo unigramas, bigramas, adjetivos e tags de parte da fala. Seus resultados experimentais sugeriram que um classificador Support Vector Machine (SVM) com características de presença unigrama supera os outros concorrentes.

Dentre as diversas abordagens supervisionadas para a análise de sentimento, na atualidade o SVM tem sido umas das mais utilizadas devido, principalmente, aos resultados promissores que tem apresentado.

A maioria dos métodos descritos a seguir são construídos sobre recursos de bag of words, e composições de sentimento são tratadas por regras manualmente.

Na abordagem de Aprendizado de Máquina (Machine Learning - ML) para a análise de sentimento, o classificador aprende automaticamente as propriedades das categorias dos documentos de treinamento pré-classificados. A classificação baseada em ML é chamada de aprendizagem supervisionada porque este processo é guiado pelo conjunto de treinamento rotulado (JAIN; SANDHU, 2015).

Neste trabalho são objetos de estudo na área de classificação os seguintes algoritmos de Machine Learning:

- Classificador Máquinas de Vetores de Suporte (SVM);
- Classificador Regressão Logística;
- Classificador Probabilístico - Algoritmo Naive Bayes (NB);
- Classificador Árvores de Decisão (DT).

5.4. Trabalhos Correlatos

Além das pesquisas voltadas para a comparação entre técnicas de aprendizado e seleção de características em Análise de Sentimentos, pode-se encontrar trabalhos que, a partir do uso das mesmas, apresentam também uma aplicação final.

Nos exemplos abaixo, destaque para trabalhos voltados para as áreas de Cidades Inteligentes e Análise de Sentimentos.

Na área de Marketing, Hopfgartner (2015) apresentou uma ferramenta de rastreamento de sentimentos no Twitter para predizer tendências de marcas de quatro empresas. Foram coletados dados de tweets em 17 cidades anglófonas (falantes da Língua Inglesa). O sistema desenvolvido classifica os tweets coletados como positivos ou negativos usando os métodos de classificação Naive Bayes com uma acurácia de 70% e Máxima Entropia com uma acurácia por volta de 60%. O sistema determina a marca vencedora de cada cidade de acordo com o percentual de tweets positivos enviados pelos usuários localizados nas cidades. As marcas vencedoras do dia podem ser monitoradas em uma página da web usando o Google Maps.

Gavilanes et al. (2019) propuseram um solução para avaliar automaticamente o engajamento de usuários nas redes sociais em diferentes países durante um evento com impacto mundial. Ao monitorar grandes eventos públicos, os usuários e mídia social podem ser fortemente influenciados por ações pontuais dos personagens participantes e pelo senso de pertencimento a um grupo específico vinculado a uma determinada área geográfica. A solução combina uma metodologia de predição de localização com uma técnica não supervisionada de análise de sentimentos.

Puri et al. (2018) apresenta um trabalho na área de Smart Governance in Smart Cities, que implica transparência no governo através do envolvimento do público. Especificamente, descreve uma pesquisa sobre ordenanças urbanas de mineração ou leis locais e as reações públicas a elas expressas no site de mídia social Twitter. São minerados decretos e tweets relacionados uns aos outros através de sua conexão mútua com as Características Inteligentes da Cidade (SCCs) e é realizada uma análise de sentimentos de tweets relevantes para analisar as opiniões do público sobre as leis locais na região urbana em questão. O processo de mineração se baseia no Conhecimento Popular (CSK), ou seja, conhecimento que é óbvio para os humanos, mas precisa ser explicitamente alimentado em máquinas para automação.

Li et al. (2016) com o objetivo de monitorar o humor dos cidadãos de uma Cidade inteligente, apresenta um artigo que usa o classificador Multinomial Naive Bayes para criar um classificador de sentimentos. Governos de todo o mundo estão tentando se aproximar de seus cidadãos para um melhor monitoramento e governança das Cidades Inteligentes. A análise de Sentimentos a partir do Twitter está abrindo novas oportunidades que podem auxiliar nesse processo.

Mankar et al. (2018) usam análise de sentimentos nos tweets coletados usando a API do Twitter e também os valores de fechamento de várias ações, buscamos construir um sistema que prevê o movimento do preço das ações de várias empresas. Tal previsão ajudaria muito um possível investidor em ações a tomar decisões informadas, o que contribuiria diretamente para seus lucros. Essas técnicas provaram ser altamente eficazes, gerando máxima precisão com investimento monetário mínimo e também economizando grandes quantidades de tempo de retornos de ações.

Hridoy et al. (2015) neste artigo apresenta a análise da informação pública a partir das mídias sociais e sua produção de resultados interessantes e insights sobre o mundo das opiniões públicas a respeito de qualquer produto, serviço ou personalidade. Os dados da rede social são um dos indicadores mais eficazes e precisos do sentimento público. O artigo apresenta uma metodologia que permite a utilização e interpretação de dados do Twitter para determinar as opiniões do público. A análise foi feita em tweets sobre o iPhone 6. A popularidade específica do recurso e a análise específica masculino-feminino foram incluídas.

Em Sluban et al. (2015) é tratado como as mídias sociais e as redes sociais contribuem para moldar o debate sobre questões sociais e políticas, mas a dinâmica desse processo não é bem compreendida. Como estudo de caso, foi monitorado a atividade do Twitter em uma ampla gama de questões ambientais. Primeiro, identificou-se os usuários e comunidades influentes por meio de uma análise de rede dos retweets. Em segundo lugar, realizou-se uma classificação baseada em conteúdo das comunidades de acordo com os principais interesses e perfil de seus usuários mais influentes. Em terceiro lugar, realizou-se uma análise de sentimentos dos tweets para identificar a inclinação de cada comunidade para um conjunto de tópicos comuns, incluindo algumas questões controversas. Essa nova combinação de análise de rede, baseada em conteúdo e de sentimento permite uma melhor caracterização dos grupos e suas tendências em redes sociais complexas.

Em Lunardi (2016) é apresentada uma tabela resumo contendo os trabalhos com classificadores baseados em multiclasse, como pode ser visualizado a seguir na Tabela 2.

Tabela 2 – Resumo dos Principais Trabalhos em Análise de Sentimento Multiclasse

Autores	Domínio	Técnica de Extração de Caracteres	Algoritmos	Acurácia (%)
Pang e Lee 2005	Filmes	Frequência de um termo	SVM One-vs-all, Regression and Metric Label	59,4 (SVM + vetor de palavras + regressão)
Goldberg e Zhou 2006	Filmes	Modelo semi-supervisionado baseado em grafos	Regressão (SVM), Metric Labeling e PSP	59,2 (regressão + PSP ou regressão)
Pak e Paroubek	Tweets	Frequência, n-gramas e POS	NB	60-80 (NB + bigramas)
Qu et al. 2010	Produtos da Amazon	Bag of opinions	Regressão	-*mostra apenas o erro quadrático médio
Long et al. 2010	Hotéis	Complexidade Kolmogorov + rede bayesiana	SVM	73,1 - 57,3 (Kolmogorov + SVM baseado em atributos)
Albornoz et al. 2011	Hotéis	Vetor de intensidade das características	Regressão Logística SVM e Árvore Funcional	46,9 (vetor + regressão)
Paltoglou e Thelwall 2013	Notícias	Palavras que expressem valência e excitação	Regressão de Vetor de Suporte, SVM (OvA)	51,8 (Excitação + SVM (OvA))

Fonte: Autor adaptada de (LUNARDI, 2016)

5.5. Considerações Finais do Capítulo

As técnicas de Análise de Sentimentos usadas nos trabalhos descritos são baseadas em dicionários léxicos ou em Aprendizagem de Máquina.

O uso de Aprendizagem de Máquina Supervisionada possibilita construir um modelo específico a partir de atributos e recursos úteis para a classificação de sentimentos em relação ao domínio. O bom desempenho do modelo depende de seleções adequadas de recursos de classificação.

Alguns desses trabalhos realizam a separação por domínio de interesse por meio de Application Programming Interface (APIs) disponibilizadas pelas Redes Sociais, ou por outros aplicativos, antes de realizar a Análise de Sentimentos. Verifica-se a necessidade que esses dados sejam representativos e com qualidade.

O Método ASCI, apresentado nesta Tese, foi desenvolvido utilizando técnicas de Aprendizagem de Máquina Supervisionada com duas Fases de Mineração de Dados. Ele difere-se dos demais trabalhos, pois aplica uma fase exclusiva de Mineração de Dados para a obtenção de dados específicos e de boa representatividade do Domínio de Interesse

Após isso, é aplicada uma Segunda Fase de Mineração de Dados, onde é realizada Análise de Sentimentos sobre o Domínio de Interesse escolhido.

A escolha do classificador que fará parte de cada Fase de Mineração de Dados é determinada durante a etapa de Treinamento. Em cada Fase de Mineração são apresentados quatro classificadores cujos desempenhos são avaliados de acordo com o cenário de domínio. O classificador que apresenta o melhor desempenho é escolhido como modelo. Esse fato não foi apresentado nos trabalhos correlatos. Alguns deles apresentam, em seus processos, apenas um ou dois classificadores.

6. PROPOSTA DO MÉTODO DE DESENVOLVIMENTO DO SISTEMA DE ANÁLISE DE SENTIMENTOS PARA O AUXÍLIO NA GESTÃO DAS CIDADES INTELIGENTES (ASCI)

Neste capítulo é apresentada a proposta desta pesquisa, um método para o Desenvolvimento de um Sistema de Análise de Sentimentos para Auxílio na Gestão das Cidades Inteligentes (ASCI).

6.1. Definição do Método de Desenvolvimento do Sistema de Análise de Sentimentos para Auxílio na Gestão das Cidades Inteligentes

O trabalho em questão propõe o desenvolvimento de um método que forneça as funcionalidades necessárias para o desenvolvimento de um sistema que seja capaz de captar informações das Redes Sociais provenientes dos cidadãos de uma Cidade Inteligente e Humana (Smart Human Cities SHC). Nesse método, as informações são captadas, pré-processadas, classificadas e filtradas de acordo com os Domínios/Classes de Interesse da Gestão das Cidades Inteligentes, e, posteriormente, processadas e analisadas quanto aos sentimentos contidos nessas informações. Finalmente, são apresentados os resultados. Portanto, este método constitui-se em uma ferramenta fornecedora de informações no auxílio da gestão das cidades inteligentes, em busca da melhoria da qualidade de vida dos cidadãos.

O método apresentado está organizado em quatro fases, similar ao processo KDD, porém, com duas fases de mineração de dados: 3ª Fase Classificação dos Domínios e 4ª Fase Análise de Sentimentos. As Fases que compõem o método de Desenvolvimento do Sistema de Análise de Sentimentos para Auxílio na Gestão das Cidades Inteligentes são apresentadas na Figura 9, sendo detalhadas na sequência.

Figura 9 – Fases do Desenvolvimento do Sistema de Análise de Sentimentos no Auxílio da Gestão de Cidades Inteligentes

Fonte: Autora

6.1.1. 1ª Fase - Captação dos Dados

As informações devem ser captadas por meio das postagens feitas pelos munícipes nas Redes Sociais, no conceito de Cidades Inteligentes e Humanas. Dessa maneira, os munícipes podem passar a ser utilizados como sensores de informações, de forma participativa e/ou indireta, por meio de seus dispositivos inteligentes. É importante ressaltar que, diferentemente dos sensores físicos que apresentam custos de instalação e manutenção, a utilização dos munícipes como sensores não necessita desses investimentos.

Inicialmente, deve-se escolher a plataforma de Rede Social da qual se deseja extrair os Dados. Para a realização da Coleta dos Dados são definidas as palavras-chave pré-estabelecidas de acordo com cada classe, onde cada classe representa um Domínio de Interesse da Gestão das Cidades Inteligentes. A definição das palavras-chave tem o objetivo de coletar o maior número de postagens referente aos domínios de interesse da Cidade Inteligente que desejam ser avaliados. No presente trabalho, é realizada uma busca com base em uma lista de palavras-

chave que aparecem no corpo do texto ou nas *Hashtags*. As *Hashtags* são palavras que designam um assunto abordado, antecidas pelo caractere cerquilha. Alguns exemplos de palavras-chave: #trânsito, #problema no trânsito, #acidente, #faltademédico, #assalto, #Segurança, #educação, etc.

As informações devem ser coletadas por meio de um script onde devem ser selecionados os parâmetros de geolocalização, contemplando apenas as postagens que estejam localizadas dentro da Cidade Inteligente e também que contemplem as palavras-chave de interesse. Todos os dados captados são armazenados em formato de arquivo texto para próxima Fase.

6.1.2. 2ª Fase - Pré-processamento dos Dados

Na mineração de texto, as resoluções para as tarefas de classificação e agrupamento devem ser precedidas por uma etapa de pré-processamento dos dados textuais, a fim de garantir maior eficácia do algoritmo de mineração (SILVA; PERES; BOSCAROLI, 2016).

Esta Fase é composta por um módulo de preparação de dados que é representado por mecanismos de pré-processamento dos dados para prepará-los para a Fase seguinte. Vale ressaltar que os textos disponibilizados nas Redes Sociais são produzidos por usuários e, em alguns casos, apresentam problemas em sua construção, como erros de ortografia, construção de fraseado errado, gíria, entre outros. Além disso, os dados devem estar preparados de acordo com as regras dos métodos de aprendizagem.

O objetivo do pré-processamento dos dados é obter um conjunto de dados (matriz) com uma representação qualificada, conforme a Figura 10, para posterior aplicação de algoritmos de mineração.

Figura 10 – Processo de Preparação da Representação dos Dados

Fonte: Autora

Após a seleção dos tweets, passam a ser realizados os seguintes processos:

- **Análise Léxica (Tokenization):** nesta etapa as palavras das mensagens são separadas em token. Por exemplo, seja a frase “Aprendendo data science no curso”. Após a tokenization, as palavras da mensagem são separadas em unidades (token) da seguinte forma: ‘Aprendendo’, ‘data’, ‘science’, ‘no’, ‘curso’. Também são retirados caracteres de pontuação, caracteres especiais, acentos, links de url e outros caracteres que não são necessários. Um exemplo de caracteres especiais na Tabela 3:

Tabela 3 - Exemplo de Caracteres Especiais

Descrição	Token
Acentos	´ ´ ~ ^
Pontuação	‘ ’ , . ; : ? !
Especiais	@ # * () &
Emoticons	:) ;) :D : (; (
HTML	 <p>

Fonte: Autora

- Case Folding: conforme o domínio, verifica-se a necessidade de remover todas as distinções de maiúsculas e minúsculas. Neste caso, ajustando todas as palavras para minúsculas;
- Lematização ou stemming é o processo de reduzir ao radical original as palavras derivadas ou flexionadas deste radical. O principal objetivo da utilização de um processo de stemming é a redução da grande dimensionalidade da representação das palavras nas aplicações de Mineração de Textos, pois, sem a remoção de prefixos e sufixos, palavras derivadas de um mesmo radical seriam consideradas como tokens distintos. Dessa forma, um único token é obtido para a representação de todas elas. São retirados o plural, gerúndio, verbos flexionados, aumentativo e diminutivo. Por exemplo, as palavras “real”, “realidade”, “realeza” e “realizado” têm sua semântica relacionada. O objetivo é a obtenção de um elemento único com uma semântica única, que represente estes elementos de texto;
- Negação: desempenha um papel vital em um sentimento e expressão de opinião. Em cada tweet, substituímos todas as negações (por exemplo, não, não, nunca, não) por “não”. A negação deve ser utilizada conforme o domínio de interesse (PANG; LEE, 2008);
- Mineração de palavras irrelevantes (stopwords): nesta etapa são retiradas as palavras consideradas com pouco ou nenhum sentimento, denominadas Stopwords. Dessa forma, a quantidade de palavras será reduzida para que possam ser usadas no treinamento apenas palavras que influenciem na determinação do sentimento final de um texto. Em geral, são retirados artigos, preposições, pronomes, numerais, conjunções, advérbios e palavras comuns para o contexto;
- Bag of Words (BOW) método de extração de características de texto: trata-se de um conjunto de palavras e a contagem da frequência de cada palavra em um documento, conforme Figura 12. Essas palavras podem ser salvas como unigramas ou bigramas a fim de contabilizar o número de ocorrências de uma dada palavra em uma classe. O unigrama corresponde a uma palavra. Palavras agrupadas em duas palavras formam bigramas, e agrupadas em n, ou seja, mais que duas, formam os n-gramas. Na Tabela 4, é possível visualizar a frase “Trânsito lento na Marginal Pinheiros” na forma de

unigramas e bigramas. Cada n-grama está separado por vírgulas e a união dos mesmos está representada pelo caractere “_”. Unigramas e bigramas são as principais formas de representação de características e possuem bons resultados na Análise de Sentimento (LIU, 2012).

Documento 1: Semáforos quebrados na avenida Morumbi e na avenida dos Bandeirantes.
Documento 2: Trânsito muito ruim na avenida dos Bandeirantes.										
	Semáforos	quebrados	na	avenida	Morumbi	e	dos	Bandeirantes	Trânsito	muito	ruim
D1	1	1	2	2	1	1	1	1	0	0	0
D2	0	0	1	1	0	0	1	1	1	1	1

Figura 11 – Bag of Words Matriz de Representação das Ocorrências das Palavras

Fonte: Autora

Tabela 4 – Exemplos de Unigrama e Bigrama

Unigrama	Trânsito, lento, na, Marginal, Pinheiros
Bigrama	Trânsito_lento, lento_na, na_Marginal, Marginal_Pinheiros

Fonte: Autora

Tendo processada a lista de n-gramas, o objetivo é obter uma representação do texto em características numéricas, no caso uma matriz M de Bag of Words. A matriz M será dada por Documentos X Ntokens, onde cada linha na matriz M contém a frequência de tokens no documento D;

- Frequência do Termo–Inverso da Frequência nos Documentos TF-IDF (Term Frequency–Inverse Document Frequency): destina-se a converter os documentos de texto em modelos vetoriais com base na ocorrência de palavras nos documentos. A fórmula TF-IDF dá a importância relativa de um termo em um corpus (lista de documentos), dada pela seguinte fórmula abaixo:
 - $TF = (\text{Número de vezes que o termo } t \text{ aparece em um documento}) / (\text{Número de termos no documento});$
 - $IDF = \log (N / n)$, onde, N é o número de documentos e n é o número de documentos em que um termo t apareceu;
 - $TF-IDF = TF * IDF$.

O valor $tf-idf$, é uma medida estatística que tem o intuito de indicar a importância de uma palavra de um documento em relação a uma coleção de documentos ou em um corpus linguístico. Ela é frequentemente utilizada como fator de ponderação (peso) na recuperação de informações e na mineração de dados. O modelo TF-IDF configura os vetores com um peso w_t para um termo t . É gerada uma representação vetorial para os textos baseada na métrica TF-IDF.

6.1.3. 3ª Fase– Mineração de Dados– Classificação dos Domínios

Nesta terceira fase ocorre a Mineração dos Dados pré-processados. O objetivo desta fase é filtrar as informações de acordo com os Domínios definidos pela Gestão das Cidades Inteligentes, por exemplo Saúde, Transporte, Educação, Abastecimento, Segurança, Meio Ambiente e outros.

A tarefa de classificação determina a geração de um mapeamento entre dados de entrada e classes (ou categorias) de saída. Se esse mapeamento é estabelecido a partir de um algoritmo de aprendizado de máquina, ou seja, raciocínio indutivo, ele demanda um conjunto de dados previamente rotulado (SILVA; PERES; BOSCARIOLI, 2016).

A técnica adotada para a realização da classificação dos dados coletados por Domínio de Interesse da Cidade Inteligente nesse trabalho é a de Aprendizagem de Máquina Supervisionada e aplicada ao Processamento de Linguagem Natural.

Primeiramente, os dados já tratados na 2ª Fase são divididos em duas partes, denominadas Treinamento e Teste. No processo de Treinamento, os algoritmos classificadores são treinados com dados rotulados para criar modelos de classificação para, posteriormente, categorizar os dados sem rótulos nos Domínios Transporte, Segurança, Educação e outras áreas de interesse.

Os dados de Treinamento são então apresentados aos classificadores duas vezes, utilizando, uma vez, a Técnica de Extração de Características BOW e, na outra vez, a Técnica de Extração de Características TF-IDF. Como resultado, são obtidos dois modelos para cada classificador.

Os desempenhos desses modelos são analisados e comparados. O classificador que apresentar melhor desempenho no Treinamento, através das métricas acurácia e f1-score, será adotado pela ferramenta para a aplicação dos dados de Teste, onde novamente será avaliado pelas mesmas métricas para comprovar sua eficácia.

Seguem os algoritmos de classificação adotados para a realização desta Fase, os mesmos foram apresentados anteriormente no Capítulo 2, na Seção 2.2.1:

- Support Vector Machine (SVM): um dos algoritmos mais utilizado entre todos os algoritmos Machine Learning. O SVM também trabalha com multiclases, onde são utilizados mais que um hiperplano. Como a saída apresenta vários domínios, para atender o propósito deste trabalho será utilizado SVC Linear (LinearSVC) para multiclases. O SVM é amplamente utilizado na classificação de textos, marketing, reconhecimento de padrões e diagnóstico médico (CHEN et al., 2015);
- Regressão Logística: nessa arquitetura é usado o modelo Logístico Multinomial (LogisticRegression) onde é possível estimar uma regressão logística onde o modelo logístico pode ser estendido quando a variável de resposta qualitativa tem mais do que duas categorias.(DAVIS; OFFORD, 1997);
- Naive Bayes: é um método de aprendizagem probabilística baseado no Teorema de Bayes. Esse algoritmo tem apresentado bons resultados em sua utilização. Para dados textuais, geralmente utiliza-se uma versão mais apropriada do algoritmo, denominado Multinomial Naive Bayes (MultinomialNB), cuja diferença está na forma em que são

computadas as probabilidades, que neste caso considera que os atributos seguem uma distribuição multinomial como acontece com as palavras nos textos;

- **Árvore de Decisão:** nessa pesquisa utiliza-se o algoritmo de classificação Random Forest (RandomForestClassifier) é um termo usado para um conjunto de árvores de decisão. Em Random Forest, temos coleções de árvores de decisão (conhecidas como “Forest”). Para classificar um novo objeto com base em atributos, cada árvore fornece uma classificação e dizemos que a árvore “vota” para essa classe. A floresta escolhe a classificação com mais votos (sobre todas as árvores da floresta).

6.1.4. 4ª Fase - Mineração de Dados - Análise de Sentimentos

A Análise de Sentimentos, também conhecida como Mineração de Opinião, é o estudo computacional das opiniões, atitudes e sentimentos das pessoas em relação a uma entidade (por exemplo, indivíduos, eventos ou temas). A classificação dos sentimentos visa prever automaticamente a polaridade do sentimento (por exemplo, positivo ou negativo) dos dados gerados pelos usuários (por exemplo, comentários, blogs, postagens e outros).

Os métodos de Classificação de Sentimentos por meio de Aprendizado de Máquina são construídos a partir de características linguísticas que dependem de um conjunto de treinamento rotulado, para aprender as características diferenciadoras dos textos, e um conjunto de testes, para avaliar o desempenho do classificador (GAVILANES et al., 2019)

Esta quarta fase consiste em, novamente utilizar a técnica de Mineração de Dados, agora com foco na Análise de Sentimentos, a partir dos conjuntos de dados que foram minerados na terceira fase, ou seja, das informações já classificadas por Domínio de Interesse.

Esses dados são divididos em duas partes, denominadas Treinamento e Teste, como na terceira Fase. No processo de Treinamento, os algoritmos classificadores são treinados com dados rotulados para classificar os Sentimentos com as seguintes polaridades: Positiva, Negativa ou Neutra.

Os dados de Treinamento são então apresentados aos classificadores duas vezes, utilizando, uma vez, a Técnica de Extração de Características BOW e, na outra vez, a Técnica de

Extração de Características TF-IDF. Como resultado, são obtidos dois modelos para cada classificador.

Os desempenhos desses modelos são analisados e comparados. O classificador que apresentar melhor desempenho no Treinamento, por meio das métricas acurácia e f1-score, será adotado pela ferramenta para a aplicação dos dados de Teste, onde novamente será avaliado pelas mesmas métricas para comprovar sua eficácia.

Os dados serão avaliados pelos algoritmos de Aprendizado de Máquina Supervisionado para realizar a técnica de Análise de Sentimentos. Serão utilizados, de maneira similar à 3ª Fase, os algoritmos de Support Vector Machine (LinearSVM), Regressão Logística (Logistic Regression), Classificação Naive Bayes (MultinomialNB) e Árvore de Decisão (RandomForest), que passam pelos mesmos processos de Treinamento e Teste.

Os resultados devem ser representados pelos valores -1, 0 e 1, que correspondem aos sentimentos Negativo, Neutro e Positivo. Por último, são apresentados gráficos informativos desses resultados, tais como: Nuvens de Palavras, Gráficos de Pizza e Gráficos de Colunas.

6.1.5. Fluxo Geral das Informações do ASCI

Na figura .12 apresenta-se o fluxo geral das informações do ASCI, conforme descrito nas seções anteriores.

Figura 12 - Método de Desenvolvimento do Sistema de Análise de Sentimentos para Auxílio na Gestão das Cidades Inteligentes

Fonte: Autora

6.2. Métricas para Avaliação do Desempenho do Classificador

Para ter um bom entendimento das métricas, faz-se necessário utilizar a Matriz de Predição, aqui denominada Matriz de Predição, que é uma matriz que apresenta os valores reais e os valores preditos do classificador.

Seja um classificador binário de duas classes (positivo e negativo) aplicado a um conjunto de teste. Para cada tupla do conjunto de teste é comparado a previsão do rótulo de classe do classificador (valor preditivo) com o rótulo de classe conhecido da tupla (valor real). O classificador pode apresentar quatro estados como resposta (HAN, JIAWEI; KAMBER, MICHELINE; PEI, 2012) :

- Verdadeiro positivo (VP): significa uma classificação correta da classe Positivo. Por exemplo, a classe real é Positivo e o modelo classificou como Positivo;
- Verdadeiro Negativo (VN): significa uma classificação correta da classe Negativo. Por exemplo, a classe real é Negativo e o modelo classificou como Negativo;
- Falso Positivo (FP): significa uma classificação errada da classe Positivo. Por exemplo, a classe real é Negativo e o modelo classificou como Positivo;
- Falso Negativo (FN): significa uma classificação errada da classe Negativo. Por exemplo, a classe real é Positivo e o modelo classificou como Negativo.

A matriz de Predição, Figura 13, é uma ferramenta útil para analisar quão bem seu classificador pode reconhecer tuplas de diferentes classes. VP e VN nos dizem quanto o classificador está acertando, enquanto FP e FN nos dizem quando o classificador está errando (HAN, JIAWEI; KAMBER, MICHELINE; PEI, 2012).

		Valor Preditivo	
		Positivo	Negativo
Valor Real	Positivo	VP	FN
	Negativo	FP	VN

Figura 13 – Matriz de Predição

Fonte: Autora

As medidas de avaliação do classificador precisão (precision), revocação (recall), medida F1(f1score) e acurácia (accuracy) apresentadas neste trabalho são descritas a seguir (HAN, JIAWEI; KAMBER, MICHELINE; PEI, 2012):

A Precisão (precision) é calculada $\text{precision} = \text{VP} / (\text{VP} + \text{FP})$, apresenta a taxa de classificação positiva correta frente aos valores preditos como positivos.

A Revocação (recall ou sensibilidade) é calculada $\text{recall} = \text{VP} / (\text{VP} + \text{FN})$, apresenta a proporção de verdadeiros positivos, predições corretas, em relação ao total de positivos da amostra.

A Medida F1 (F1-score) é calculada como :

$$\text{F1-score} = 2 * ((\text{precision} * \text{recall}) / (\text{precision} + \text{recall}))$$

Essa medida é a média harmônica entre a precisão (precision) e a revocação (recall). O resultado é um indicativo, quanto mais próximo de 1, melhor é o desempenho do classificador e resultados próximos de 0 o desempenho é ruim para a classe atrelada ao conceito de positivo.

A Acurácia (Accuracy) é calculada $\text{Accuracy} = (\text{VP} + \text{VN}) / (\text{VP} + \text{FP} + \text{VN} + \text{FN})$, apresenta a medida de desempenho global, é a proporção de classificações corretas, tanto de casos positivos quanto negativos sobre o total de exemplos.

No caso de Matriz de Predição do tipo Multiclasse, Figura 14, ou seja, com três ou mais classes, as métricas usadas são as mesmas da classificação binária. A métrica é calculada para cada classe, tratando-a como um problema de classificação binária após agrupar todas as outras classes como pertencentes à segunda classe. Em seguida, a média da métrica binária é medida em todas as classes para obter uma métrica de média. Na Figura 14, os valores em destacados pelos blocos em azul, na diagonal, são os valores previstos pelo classificador de maneira correta para cada classe.

		Valor Preditivo		
		C1	C2	C3
Valor Real	C1	125	11	2
	C2	0	285	5
	C3	26	3	44

Figura 14 –Matriz de Predição Multiclasse

Fonte: Autora

As métricas adotadas neste trabalho são a acurácia e o f1-score.

6.3. Uso do ASCI na Gestão das Cidades Inteligentes

O Sistema desenvolvido pelo Método ASCI pode realizar a Análise de Sentimentos da população de uma Cidade Inteligente referente aos Domínios de Interesse dos Governantes e de seus munícipes. Como resultado, são apresentadas informações que permitem que os gestores dessas cidades possam tomar as melhores decisões de acordo com cada cenário.

Após a Análise de Sentimentos das informações, as mesmas recebem uma inferência de polaridade (Negativa, Neutra ou Positiva). É importante ressaltar que essa Análise de Sentimentos realizada nos Domínios de Interesse pode apresentar diferentes níveis de granularidade, dependendo da definição do conjunto de dados a serem estudados.

O operador da central pode visualizar essas informações por meio de gráficos disponíveis em um módulo de apresentação dos resultados do Sistema de Gestão com foco na Análise de Sentimentos.

Ao avaliar os sentimentos da população é possível prever situações críticas, negativas, entre os usuários, onde ações podem ser tomadas em busca de melhorias na qualidade vida da população. Essa análise também pode ser feita por meio de sensores e nuvens de palavras, ambas de fácil interpretação visual.

7. ESTUDO DE CASO - SISTEMA DE GESTÃO DAS CIDADES INTELIGENTES UTILIZANDO ANÁLISE DE SENTIMENTOS

Como caso de estudo, é apresentado um estudo de caso no domínio do transporte, em uma determinada região da cidade de São Paulo. O resultado gerado a partir deste estudo de caso tem por objetivo contribuir com o Sistema de Transporte Inteligente da Cidade, provando que o uso de Análise de Sentimento pode ser uma ferramenta importante a mais para auxiliar os gestores da cidade e também sua população.

Apesar de ter apresentado avanços na questão da mobilidade urbana nos últimos anos, como a extensão dos corredores de ônibus, aumento no número de estações de metrô, e a instalação de ciclovias, São Paulo ainda não conseguiu atingir um nível satisfatório no sistema de transporte urbano.

É importante citar que, neste trabalho, a cidade de São Paulo, embora não possua todas as características das chamadas “Cidades Inteligentes”, será considerada como sendo uma delas.

Para estudo de caso foi utilizada a CET Companhia de Engenharia de Tráfego e o uso de Análise de Sentimentos no Auxílio da Gestão do Tráfego numa cidade Inteligente.

7.1. Estudo de Caso: CET - Companhia de Engenharia de Tráfego

Companhia de Engenharia de Tráfego - CET é uma empresa de economia mista vinculada à Prefeitura de São Paulo, responsável pelo gerenciamento, operação e fiscalização do sistema viário da cidade.

Foi criada em 1976 cuja missão é contribuir para a Mobilidade da cidade de São Paulo, por meio do estímulo e da integração dos diferentes modais de deslocamentos, primando pela segurança viária, visando a melhoria da qualidade de vida (CET, 2018).

Atualmente a CET é composta por oito Gerências de Engenharia de Tráfego, conhecidas como GETs, atuando em 1.523 Km² de área na cidade onde há 17.291 Km de vias, apresentadas na Figura 15. Em São Paulo, segundo estimativa do IBGE (Instituto Brasileiro de Geografia e Estatística), existem cerca de 12,2 milhões de habitantes.

Até julho de 2018, o DETRAN (Departamento Estadual de Trânsito de São Paulo) possuía em seus registros cerca de 8,8 milhões de veículos.

Diariamente, são realizadas 29,7 milhões de viagens no modo motorizado e 14 milhões de viagens no modo não motorizado (sendo 13,7 milhões de viagens a pé e 333 mil viagens de bicicleta).

Figura 15 – Mapa do Município de São Paulo subdividido por área de atuação das GETs

Fonte: (CET, 2018)

As principais áreas de atuação da CET na cidade de São Paulo são: Operação de Tráfego, Fiscalização, Projeto/Planejamento, Sinalização, Segurança no Trânsito e Educação.

Atividades realizadas diariamente:

- Sistema Viário Monitorado: 868 Km dos principais corredores;
- 143 ROTAS (turnos de pico) percorrendo 1.749 km de vias;
- 17 VIAS E FAIXAS REVERSÍVEIS (37,44 km) (Veículos e Transporte Coletivo);
- 244 PONTOS OPERACIONAIS ativados nos turnos de pico;
- 102 GUINCHAMENTOS para remover veículos estacionados irregularmente;

- 3 ACOMPANHEMENTOS DE MANIFESTAÇÕES;
- 53 ACIDENTES atendidos. São 16 acidentes com vítimas. Dados de registros policiais - 02 vítimas fatais/dia (1 motociclista).

Também são realizadas Operações Especiais de acordo com a demanda: carnaval de rua, virada cultural, eventos religiosos; fórmula1, stok car, fórmula truck; parada lgbt; São Silvestre; reveillon; corridas de ciclistas e pedestres; operação em estádios de futebol, obras, manutenção como de pinturas das faixas de sinalização; transporte de grandes cargas; dentre outros.

Para atender a essa demanda, fazem parte dos recursos humanos 2.046 agentes municipalizados (marronzinhos) e 298 Operadores de Monitoramento de Informações de Trânsito (OMITs), que trabalham na Central. Alguns dos principais recursos materiais são: 857 veículos (pick-ups, van, motos, guinchos, caminhão e ônibus), 502 câmeras de monitoramento e 75 Painéis de Mensagens Variáveis (PMVs) (fonte CET outubro/2018).

A CET conta com diversos canais de comunicação entre a população e a companhia, além de parcerias estratégicas com Corpo de Bombeiro, Polícia Militar, Rádio Trânsito, Waze, Google e outros.

7.1.1. Central de Operações

Fazem parte da Central de Operações os seguintes canais de informações, Figura 16:

Monitoramento de Câmeras nas principais vias, neste caso os operadores monitoram as câmeras e assim que percebem alguma intercorrência a informam através de rádio ou smartphones às viaturas e/ou corpo de bombeiros para que sejam tomadas as devidas providências.

De forma similar às câmeras, os PACs (Posto Avançado de Campo), situados em oito pontos estratégicos da cidade, como no último andar de edifícios, observam todas as ocorrências e, no momento que elas acontecem, os mesmos avisam a central via rádio ou smartphone e os OMITs acionam as viaturas e/ou corpo de bombeiros.

Telefone do Trânsito 1188 (24 horas), onde os munícipes entram em contato com a CET informando problemas referentes ao trânsito. Os operadores recebem a informação que prontamente é cadastrada e em seguida, via sistema, são enviadas para os OMITs de uma determinada GET, que conforme a localização e o problema acionam as viaturas e/ou corpo de bombeiros através de smartphones ou rádios, para as devidas providências.

De forma similar, tweets são recebidos informando problemas referente ao trânsito. Eles são avaliados e conforme a intercorrência os mesmos são informados via sistema para os OMITs de uma determinada GET, que conforme a localização e problema acionam as viaturas e/ou corpo de bombeiros através de smartphone ou rádios, para as devidas providências.

Figura 16 - Central de Operações - CET

Fonte: (CET, 2018)

Os painéis de monitoramento de Lentidão de Tráfego são alimentados pelos operadores por meio das informações recebidas dos operadores em campo. O gráfico (Figura 17) e o mapa (Figura 18), são atualizados a cada 30 minutos e essas informações são distribuídas para os meios de comunicação, como rádio e TV.

Figura 17 – Gráfico de Lentidão na Cidade de São Paulo

Fonte: (CET, 2018)

Figura 18 – Mapa de Lentidão na Cidade de São Paulo

Fonte: (CET, 2018)

7.1.2. Ocorrências no Trânsito

O Tempo médio de deslocamento para atendimento após notificação de um acidente é em média de 9,4 minutos. No caso de remoção chega a 16.9 minutos. Ou seja, o tempo total de um atendimento completo é de aproximadamente 27 minutos.

Os acidentes ocorridos no decorrer do dia geram custos para o município nas mais diversas áreas, tais como:

- Meio ambiente, retirada de resíduos tóxicos ocasionados pelos acidentes;
- Saúde, gastos com ambulância e hospital, onde a vítima de trânsito ocupa um lugar de alguém que está com alguma enfermidade;
- Sequelas por causa do acidente;
- Em geral as vítimas estão na faixa de idade entre 20 e 30 anos, fato que acaba muitas vezes ocasionando aposentadoria por invalidez em pessoas jovens.

Em busca de um melhor monitoramento e com o avanço tecnológico, as câmeras entraram com intuito de substituir os PACs (Posto Avançado de Campo) e oferecer um melhor serviço de monitoramento.

Apesar das câmeras captarem imagens em 360° de forma instantânea, as mesmas ainda não oferecem uma melhor percepção das ocorrências quando comparadas com os PACs.

A ação humana faz uma diferença enorme apesar de toda tecnologia. Para a CET, uma das ferramentas de comunicação de grande importância é o ser humano (fonte Ricardo Pires Gerente da Central CET 2018).

7.2. Implementação do Estudo de caso Sistema de Análise de Sentimento para Auxílio na Gestão das Cidades Inteligentes - ASCI

Para a implementação deste estudo de caso (ASCI) desenvolvido neste trabalho utilizou-se a linguagem de programação Python versão 3.6, por meio do ambiente de desenvolvimento integrado Anaconda Navigator versão 1.9.6. Foi adotada esta linguagem devido a suas características, tais como: código aberto, alto nível, desenvolvimento rápido, ampla adoção na indústria de desenvolvimento de software (Google e YouTube) e também no ambiente

científico. Como resultado, a linguagem disponibiliza bibliotecas para Processamento de Linguagem Natural (PLN) e Aprendizado de Máquina (ML).

O estudo de caso desenvolvido realiza o monitoramento automático dos sentimentos expressos pelas pessoas, a respeito do trânsito da Cidade de São Paulo dentro da Rede Social Twitter. As informações postadas pela população são captadas utilizando o conceito Smart Human Cities (SHC) adotado neste trabalho, no qual os cidadãos colaboram com as informações no auxílio da gestão das cidades. Essas informações são processadas e analisadas, por meio da Análise de Sentimentos para posteriormente os resultados serem apresentados.

7.2.1. 1ª Fase realização da Captação dos Dados

A Rede Social escolhida para este Estudo de Caso foi o Twitter, uma Rede Social em formato de microblog na qual os usuários compartilham opiniões sobre uma variedade de tópicos e expressam seus sentimentos por meio de mensagens curtas, limitadas a 280 caracteres, sobre os mais diversos assuntos. Dessa forma, o Twitter é utilizado como ferramenta provedora de dados para esta pesquisa.

O Twitter já provou seu valor e potencial em domínios que vão desde detecção de notícias até detecção de tráfego em tempo real. Sua comunidade está crescendo continuamente e, atualmente, o número de usuários ativos é de cerca de 313 milhões, registrando uma média diária de 400 milhões de novos posts. (PEREIRA, 2017).

Para a preparação dos dados desta pesquisa, foram coletados cerca de 800 tweets em dias alternados, nos meses de março e abril de 2019, por meio da API do Twitter, utilizando palavras-chave com foco nas áreas de transporte, segurança e educação, temas que fazem parte do domínio de interesse do método. Também foi usado o parâmetro de geolocalização para selecionar uma determinada região da cidade de São Paulo, Figura 19, por meio de coordenadas (latitude, longitude e raio). Outro parâmetro é a escolha do idioma através do parâmetro lang. Seguem abaixo as palavras-chaves, o idioma e a geolocalização, para as três áreas alvo (Anexo1):

- Palavras-chave tema Transporte: 'CETSP_', CET, trânsito, acidente, colisão, engarrafamento, batida de carro, engavetamento, atropelamento, capotamento,

semáforo, farol quebrado, trânsito bom, trânsito livre, tráfego normal, trânsito normal, acidente com vítima, buraco na rua, interditada no sentido, motoqueiro, faixa bloqueada, trânsito intenso;

- Palavras-chave tema Educação: 'professor, professora, férias escolares, merenda escolar, faculdade, universidade, diretor da escola, diretora da escola, aulas, aula, educação, MEC,colégio, escola, aprendizado em sala de aula, rede estadual de ensino, secretaria da educação, ensino fundamental, ensino médio ;
- Palavras-chave tema Segurança: 'traficante, tráfico de drogas, tiro, latrocínio, assassinato, pedofilia, crime, assalto, roubo, bullying, PM, polícia, estupro, tiro, tiroteio, violência, briga, torcida organizada, sequestro, meliante, arma, armado, ronda, assassinado, fuzilado, atentado, terrorista, segurança pública, polícia militar;
- Língua: lang= pt;
- Localização, conforme Figura 19: geocode=-23.533773,-46.625290, raio de 10km.

Figura 19 – Mapa do Centro da Cidade de São Paulo

Fonte: (VIVIDPLANET, 2019)

Os tweets captados contêm informações, como nome de usuário, ID do tweet, texto, data e hora da criação, etc. Nesta pesquisa para a modelagem dos classificadores, inicialmente as informações consideradas importantes são ID e o texto.

Para obter uma classificação de sentimento com alta precisão, as técnicas de Aprendizado Supervisionado exigem uma grande quantidade de dados **rotulados manualmente** (ZHOU; ZHAO; ZENG, 2014).

Devido a especificidade das informações coletada no Twitter, neste trabalho os 800 tweets foram avaliados e rotulados de forma manual por três pessoas nos temas Transporte, Segurança e Educação. Como resultado foram separados 543 tweets formando três Corpus, distribuídos nos seguintes temas: 269 em Transporte, 137 em Segurança e 137 em Educação.

Devido às particularidades dos textos que se apresentam com linguagem coloquial e possuem características específicas da cidade de São Paulo de acordo com o domínio (Transporte, Segurança e Educação), fez-se necessário que os dados captados sejam rotulados manualmente para poder treinar os classificadores de acordo com a funcionalidade esperada.

Ao analisar os textos dos tweets por domínio classificou-se como resposta os valores: 0, 1 e 2, onde 0 significa que o tweet é referente a transporte, 1 referente a segurança e 2 referente a educação.

O corpus referente a Transporte passou por uma segunda avaliação manual referente à Análise de Sentimento, sendo os tweets foram identificados um a um como positivo ou neutro ou negativo.

Na avaliação do corpus de Transporte para Análise de Sentimentos espera-se como resposta os valores: -1, 0 e 1, onde “-1” significa que o texto do tweet é negativo, “0” significa que o texto do tweet é neutro e 1 significa que o texto do tweet é positivo.

Os tweets que não foram utilizados nesse processo foram descartados devido ao fato de estarem repetidos ou por se tratarem de assuntos referentes a outras temáticas que não fazem parte do escopo deste trabalho.

7.2.2. 2ª Fase: realização do Pré-processamento dos Dados

O pré-processamento dos dados de texto é uma etapa essencial para tornar os dados prontos para mineração. O objetivo desta fase é limpar os ruídos que são menos relevantes, como pontuação, caracteres especiais, números e termos que não carregam muita ponderação no

contexto do texto. Nesta fase todos os corpus foram unidos para a realização dos seguintes procedimentos:

- Remoção dos caracteres @user do Twitter;
- Remoção de Pontuações, Números e Caracteres Especiais (...https://t.co/[\\w*])([\\-|/?!,:;()\\]);
- Remoção de palavras curtas.

O resultado desses procedimentos pode ser visualizados na Figura 20.

id	dominio	tweet	tidy_tweet	
0	2	0	trânsito ruim fui de bicicleta e cheguei na hora!	trânsito ruim fui bicicleta cheguei hora
1	3	0	Cheguei rapidinho em casa o trânsito na marginal estava bom!	cheguei rapidinho em casa trânsito margin estava bom
2	4	0	@CETSP_ semáforo parado na Salim Farah Maluf com a Celso Garcia trânsito mega!!!!!!	semáforo parado salim farah maluf com celso garcia trânsito mega
3	5	0	@fvguima Fora que o problema não é só a enchente em si mas a total falta de preparo pra lidar com transporte em di... https://t.co/Yq2uDWbh7b	fora que problema não enchent total falta preparo pra lidar com transport
4	6	0	Chuva semáforo quebrado no cruzamento da Dr Abrahão Ribeiro com a Rua Norma trânsito. Essa é São Paulo. Cadê a... https://t.co/QQ3OGqWnLh	chuva semáforo quebrado cruzamento abrahão ribeiro com rua norma trânsito essa são paulo cadê
5	7	0	@CETSP_ Semáforo quebrado desde domingo dia 10/03 nas ruas Ouvidor Peleja e Afonso Celso. Vila Mariana São Paulo... https://t.co/upMhmEGE5q	semáforo quebrado desde domingo dia rua ouvidor peleja afonso celso vila mariana são paulo
6	8	0	@ROTA Eu prefiro o maps pq quando você traça rota nele ele automaticamente joga pro caminho mais rapido (sem... https://t.co/gP8l6nPltY	prefiro map quando você traça rota nele ele automaticament joga pro caminho mai rapido
7	9	0	Estou Presa Num Engarrafamento \nSem Facebook Sem Whats Sem Instagram \nEntao Me Aguentem Tuitando no Transito	estou presa num engarrafamento \nsem facebook sem what sem instagram \nentao aguentem tuitando transito
8	10	0	@UOLNoticias CET companhia de engarrafamento de trânsito.	cet companhia engarrafamento trânsito
9	11	0	a primeira -e acho que única vez- que eu vi alguém morto foi num acidente na av tiradentes era um motoqueiro esti... https://t.co/Zd6lFRQj7	primeira acho que única vez que alguém morto foi num accident tiradent era motoqueiro esti

Figura 20 – Remoção de Pontuações, Números, Caracteres Especiais e Palavras com até 2 Letras

Fonte: Autora

Em seguida é realizada a tokenização, onde as cadeias de texto são divididas em tokens, que são termos ou palavras individuais. Também é realizado o processo de lematização ou stemming, onde são removidos os sufixos das palavras derivadas de um mesmo radical. Conforme Figura 21.

```

0                                     [trânsito, ruim, fui, bicicleta, cheguei, hora]
1                                     [cheguei, rapidinho, em casa, trânsito, margin, estava, bom]
2                                     [semáforo, parado, salim, farah, maluf, com, celso, garcia, trânsito, mega]
3                                     [fora, que, problema, não, enchent, total, falta, preparo, pra, lidar, com, transport]
4 [chuva, semáforo, quebrado, cruzamento, abrahão, ribeiro, com, rua, norma, trânsito, essa, são, paulo, cadê]
5 [semáforo, quebrado, desde, domingo, dia, rua, ouvidor, peleja, afonso, celso, vila, mariana, são, paulo]
6 [prefiro, map, quando, você, traça, rota, nele, ele, automatica, joga, pro, caminho, mai, rapido]
7 [estou, presa, num, engarrafamento, \nsem, facebook, sem, what, sem, instagram, \nentao, aguentem, tuitando, transito]
8                                     [cet, companhia, engarrafamento, trânsito]
9 [primeira, acho, que, única, vez, que, alguém, morto, foi, num, acid, tirad, era, motoqueiro, est]
Name: tidy_tweet, dtype: object

```

Figura 21 – Tokenização e Stemming dos textos

Fonte: Autora

7.2.3. 3ª Fase - Mineração de Dados, Classificação dos Domínios Transporte, Segurança e Educação

a) Treinamento dos Algoritmos

O Corpus representado por um total de 470 tweets (196 no Domínio Transporte, 137 Segurança e 137 Educação) é dividido em duas partes, sendo 70% (329 tweets) para Treinamento e Validação do Treinamento e 30% (141 tweets) para testes (verificação da adequação das técnicas empregadas), Figura 25. Esses dados são implementados pelos classificadores com o intuito de identificar o classificador com melhor desempenho para ser adotado pela ferramenta. A distribuição dos dados das amostras de Treinamento e Teste pode ser visualizada na Figura 26.

Figura 25 – Distribuição do Corpus de Tweets para Treinamento e Teste

Fonte: Autora

Figura 26 – Gráfico de Distribuição de Todo o Conjunto de Tweets

Fonte: Autora

Os algoritmos de Classificação implementados neste trabalho utilizam as técnicas de extração de características BOW e TD-IDF para extração de características dos textos preparadas na fase anterior.

A Validação Cruzada (Cross Validation) utilizada é o método k-fold, que consiste em dividir o conjunto total do corpus em k subconjuntos mutuamente exclusivos do mesmo tamanho. Um dos subconjuntos é separado para teste e os restantes, no caso k-1, são utilizados para estimação dos parâmetros e, posteriormente, é calculada a acurácia do modelo. Esse processo é realizado k vezes alternando de forma circular o subconjunto de teste. A Figura 27 mostra o esquema realizado pelo k-fold (HAN, JIAWEI; KAMBER, MICHELINE; PEI, 2012).

Realiza-se um estudo de Validação Cruzada com quatro algoritmos de classificação aplicados no Corpus de Estudo. Dessa maneira, são obtidos os desempenhos dos classificadores por meio da Acurácia Média (Figuras 28 e 29) da classificação da amostra de tweets nos três domínios. O objetivo foi realizar testes com os classificadores, verificando a qualidade em prever a classificação para um ou mais novos exemplos de dados no domínio de interesse:

1. SVC Linear (LinearSVC) para multiclass;
2. Regressão Logística Multinomial (LogisticRegression);
3. Multinomial Naive Bayes (MultinomialNB);
4. Random Forest (RandomForestClassifier).

Figura 27 – Validação Cruzada – Método k-fold

Fonte: Autora

Figura 28 – Boxplot da Acurácia da Classificação nos três Domínios – Técnica BOW

Fonte: Autora

Figura 29 – Boxplot da Acurácia da Classificação nos três Domínios – Técnica TF-IDF

Fonte: Autora

A partir desses gráficos, verifica-se dentre os quatro algoritmos, os que apresentam melhor desempenho em ambos os modelos (BOW e TF-IDF) são LinearSVC e LogisticRegression, pois os valores obtidos para a acurácia foram maiores.

A seguir, os classificadores são implementados com as técnicas de extração de características de texto (BOW e TF-IDF) e os seus resultados referente às métricas precision, recall, f1-score e acurácia são apresentados juntamente com a matriz de Predição. Todas essas métricas foram definidas no Capítulo 5, na Seção 5.2.6.

Observa-se na Figura 30 que o classificador LinearSVC obteve um melhor desempenho com a técnica de extração BOW, com uma acurácia de 90,9% e um f1-score de 90%.

Também observa-se na Figura 31 que o classificador Logist Regression obteve um melhor desempenho com a técnica de extração TF-IDF, com uma acurácia de 90,9% e um f1-score de 91%.

Figura 30 – Desempenho do algoritmo LinearSVC – Treinamento nos Domínios

Fonte: Autora

Figura 31 – Desempenho do algoritmo Logistic Regression - Treinamento nos Domínios

Fonte: Autora

Observa-se na Figura 32 que o classificador Multinomial Naive Bayes obteve melhor desempenho com a técnica de extração TF-IDF, com uma acurácia de 88,9% e um f1-score de 88%.

Figura 32 – Desempenho do algoritmo Multinomial Naive Bayes - Treinamento nos Domínios

Fonte: Autora

Observa-se na Figura 33 que o classificador Random Forest obteve melhor desempenho com a técnica de extração BOW, com uma acurácia de 85,9% e um f1-score de 84%.

Figura 33 – Desempenho do algoritmo Random Forest - Treinamento nos Domínios

Fonte: Autora

b) Validação do Treinamento

Após o treinamento, verifica-se que o classificador que obteve melhor desempenho foi o Logistic Regression com Extração de Característica TF-IDF, que obteve acurácia de 90,9% e f1-score de 91%.

Na Figura 34 são apresentados, os resultados dos tweets aplicados no classificador Logistic Regression para validação do Treinamento, comparando-se os dados Reais e dados Preditos por Domínios, tendo sido computados os resultados de 99 amostras.

Figura 34 –Validação do Treinamento – Classificação dos Tweets por Domínios

Fonte: Autora

Na Figura 35 visualizam-se também os resultados da Classificação por Domínio obtidos na Validação do Treinamento, com os Dados Reais e Dados Preditos.

Figura 35 – Gráficos Validação do Treinamento – Classificação dosTweets por Domínio

Fonte: Autora

c) Resultados Obtidos no Corpus Teste

Numa terceira etapa, apresenta-se ao classificador Logistic Regression o Corpus de Teste (141 tweets), obtendo-se como resultado uma acurácia de 93,6% e um f1-score de 93%, conforme apresentado na Figura 36:

Figura 36 – Desempenho do classificador Logistic Regression nos Domínios - Corpus Teste

Fonte: Autora

A seguir, são apresentadas as Figuras 37 e 38 mostrando os resultados obtidos com os tweets previamente separados para tal finalidade (141 tweets). Na Figura 37, é apresentado o resultado dos tweets do Corpus Teste, comparando-se dados Reais e dados Preditos por Domínios.

Figura 37 – Classificação dos Tweets por Domínios – Corpus Teste

Fonte: Autora

Na Figura 38, visualizam-se também os resultados da classificação por Domínio obtidos no Teste, com os Dados Reais e Dados Preditos.

Figura 38 – Gráficos Classificação dos Tweets por Domínios – Corpus Teste

Fonte: Autora

Na Figura 39, são apresentadas algumas amostras dos Tweets do Corpus Teste contendo os valores Reais e Preditos, representados pelas colunas “domínio” e “pred”, onde, por exemplo, na linha, com id 330, verifica-se que a predição coincidiu com o valor predito, o que não ocorreu com a linha com id 331.

id	dominio	tweet	pred	
0	330	0	Já nos fins de semana a interdição acontece das 16h de sábado até o meio-dia de domingo. Já na direção do bairro a... https://t.co/bkGSrnfBgl	0
1	331	1	RT @MarciaVicencia: @ProfBunazar @DaniloGentili Medão de sair de casa principalmente a noite! Medão!!! Arma a mim de nada adiantaria ne	2
2	332	0	Operação tapa-buraco na avenida Interlagos na #ZS sentido bairro na altura da avenida Sargento Geraldo Santana... https://t.co/1la6wZ4XgZ	0
3	333	0	Precisa encontrar um terminal de ônibus ou saber os horários e linhas que operam nele? Acesse: https://t.co/o5jbkM9rYv	0
4	334	0	Também pode ser implantada uma faixa reversível na Rua Bento Branco de Andrade Filho sentido João Dias. https://t.co/LpYkqwPmNW	0
5	335	0	Vai curtir o @LollapaloozaBr? Uma dica é pegar a linha 606F/10 Circular Lollapalooza - Autódromo ao lado da Est... https://t.co/IS0DmEBmqw	0
6	336	0	Ouvintes relatam um engarrafamento na pista lateral da #RadialLeste na altura do viaduto Bresser (radiotransitofm)	0
7	337	0	Domingo (7) obras de modernização alteram a circulação dos trens na Linha 3-Vermelha. No trecho entre Itaquer... https://t.co/By850UZqcl	0
8	338	0	Ali pertinho do viaduto há um importante local de onde saem linhas para várias regiões da cidade. Você saberia dizer qual é?	0
9	339	0	A partir das 23h30 há previsão de bloqueio no Túnel Jânio Quadros para limpeza. A interdição deve ocorrer no senti... https://t.co/8ME9jxk6YQ	0
10	340	1	Minha intuição é que os publishers não vão se salvar os artistas não vão vender CDs e DVDs e a pirataria vai enco... https://t.co/NNJWl2maJS	1
11	341	2	um professor meu do ensino médio filiado ao PCdoB falando pra alunos parafraseadamente que ele achava que o déf... https://t.co/VuEmAcX0sM	2
12	342	0	queria saber o que aconteceu na Rebouças pra estar esse trânsito o ônibus não sai do lugar	0
13	343	1	RT @kepler_johann: @uisnassif São animais que deveriam ser presos não fosse a polícia paulista sempre ao lado dos opressores. Esses caras	1
14	344	0	Se você vai para o Lollapalooza utilize a Linha 9-Esmeralda da @CPTM_oficial e fique atento ao horário de funciona... https://t.co/wOPLJl6kbO	0

Figura 39 –Tweets por Domínios - Corpus Teste

Fonte: Autora

A Figura 40 apresenta as principais Hashtags dessa amostra.

Figura 40 Hashtags da Amostra Tweets por Domínio – Corpus Teste

Fonte: Autora

7.2.4. 4ª Fase – Mineração de Dados - Análise de Sentimentos

a) Treinamento dos Algoritmos

Nessa etapa, utiliza-se o Corpus de tweets do Domínio Transporte representado por um total de 260 tweets, sendo 210 tweets para Treinamento e Validação e 50 tweets para Teste. Vale observar que desse conjunto de 260 tweets, a maioria foi constituída pelo mesmo conjunto de tweets da 3ª fase, tendo sido acrescentados alguns outros.

A distribuição dos dados das amostras de Treinamento e Teste pode ser visualizada na Figura 41.

Figura 41 – Gráfico de Distribuição do conjunto de Tweets, Treinamento e Teste, do Domínio Transporte
Fonte: Autora

Os algoritmos de Classificação utilizados nesta fase são os mesmos descritos na 3ª Fase.

Seguem os gráficos com as acurácias médias obtidas por cada classificador por meio do método Validação Cruzada, nas Figuras 42 e 43.

Figura 42 – Boxplot da Acurácia da Classificação dos Sentimentos no Domínio Transporte – Técnica BOW
Fonte: Autora

**Figura 43 – Boxplot da Acurácia da Classificação dos Sentimentos
no Domínio Transporte – Técnica TF-IDF**

Fonte: Autora

A partir desses gráficos, verifica-se a acurácia média dos classificadores. Dentre os quatro algoritmos, os que apresentam melhor desempenho em ambos os modelos (BOW e TF-IDF) são LinearSVC e LogisticRegression.

A seguir, os classificadores são implementados e os seus resultados apresentados, como verificado nas Figuras 44, 45, 46 e 47.

Observa-se na Figura 44 que o classificador LinearSVC obteve melhor desempenho com a técnica de extração BOW, com uma acurácia de 85,7% e um f1-score de 85%.

**Figura 44 – Desempenho do algoritmo LinearSVC – Treinamento dos Sentimentos
no Domínio Transporte**
Fonte: Autora

Observa-se na Figura 45 que o classificador Logist Regression obteve os mesmos resultados com as técnicas de extração BOW e TF-IDF, com uma acurácia de 84,1% e um f1-score de 83%. Apesar disso, a técnica de extração BOW pode ser considerado melhor, pois o f1-score obteve melhores resultados individuais em dois dos três sentimentos.

Observa-se na Figura 46 que o classificador Multinomial Naive Bayes obteve melhor desempenho com a técnica de extração TF-IDF, com uma acurácia de 82,5% e um f1-score de 82%.

Também observa-se na Figura 47 que o classificador Random Forest obteve melhor desempenho com a técnica de extração BOW, com uma acurácia de 74,6% e um f1-score de 71%.

Figura 45 – Desempenho do algoritmo Logistic Regression - Treinamento dos Sentimentos no Domínio Transporte
Fonte: Autora

Figura 46 – Desempenho do algoritmo Multinomial Naive Bayes – Análise de Sentimentos no Domínio Transporte
Fonte: Autora

**Figura 47 – Desempenho do algoritmo Random Forest - Treinamento dos Sentimentos
no Domínio Transporte**
Fonte: Autora

b) Validação dos Resultados

Após o treinamento escolheu-se o classificador LinearSVC, com Extração de Característica BOW, que obteve melhor desempenho quanto à acurácia, com 85,7% e quanto ao f1-Score, com 85%.

Na Figura 48, são apresentados, em formato de gráfico de barras, os resultados dos tweets aplicados no classificador LinearSVC para validação do Treinamento, comparando-se os dados Reais e dados Preditos por tipo de Sentimento.

Figura 48 – Validação do Treinamento – Classificação dos Tweets por tipo de Sentimento

Fonte: Autora

Na Figura 49, também visualiza-se a classificação por Tipo de Sentimento obtida na Validação do Treinamento, com os Dados Reais e Dados Preditos.

Figura 49 – Gráficos Validação doTreinamento – Classificação dos Tweets por tipo de Sentimento

Fonte: Autora

c) Resultados Obtidos no Corpus Teste

Numa segunda etapa, apresenta-se o Corpus de Teste (50 tweets) ao classificador LinearSVC para a Análise dos Sentimentos. Obteve-se como resultado uma acurácia de 84% e um f1-score de 83%, conforme apresentado na Figura 50:

**Figura 50 – Desempenho do classificador LinearSVC – Sentimentos no Domínio Transporte –
Corpus Teste**

Fonte: Autora

Na Figura 51, é apresentado o resultado dos tweets do Corpus Teste classificados em um gráfico comparativo entre dados Reais e dados Preditos por Tipo de Sentimento.

**Figura 51 – Classificação dos Tweets por Sentimentos
no Domínio Transporte – Corpus Teste**

**Figura 54 –Nuvem de Palavras com Sentimentos Negativo das Informações dos Tweets
no Domínio do Transporte**

Fonte: Autora

**Figura 55 –Nuvem de Palavras com Sentimentos Neutro das Informações dos Tweets
no Domínio do Transporte**

Fonte: Autora

Na Figura 56, são apresentadas amostras dos Tweets do Corpus Teste contendo os valores Reais e Preditos, representados pelas colunas “label” e “pred”.

	id	label	Tweet	pred
0	330	0	Já nos fins de semana a interdição acontece das 16h de sábado até o meio-dia de domingo. Já na direção do bairro a... https://t.co/bkGSmsBgl	1
1	333	0	Precisa encontrar um terminal de ônibus ou saber os horários e linhas que operam nele? Acesse: https://t.co/o5jkm9rYv	0
2	335	0	Vai curtir o @LollapaloozaBr? Uma dica é pegar a linha 606F/10 Circular Lollapalooza - Autódromo ao lado da Est... https://t.co/AS0DmEBmqw	0
3	336	-1	Ouvintes relatam dois veículos quebrados pela Radial Leste sentido centro na altura da estação Vila Matilde do Me... https://t.co/AkPJvk5Atr	-1
4	337	0	Domingo (7) obras de modernização alteram a circulação dos trens na Linha 3-Vermelha. No trecho entre Itaquer... https://t.co/By850UZqcl	0
5	338	0	Ali pertinho do viaduto há um importante local de onde saem linhas para várias regiões da cidade. Você saberia dizer qual é?	0
6	339	0	A partir das 23h30 há previsão de bloqueio no Túnel Jânio Quadros para limpeza. A interdição deve ocorrer no senti... https://t.co/8ME9jx6YQ	0
7	342	-1	queria saber o que aconteceu na Rebouças pra estar esse trânsito o ônibus não sai do lugar	-1
8	344	0	Se você vai para o Lollapalooza utilize a Linha 9-Esmeralda da @CPTM_oficial e fique atento ao horário de funciona... https://t.co/wOPLJl6kbO	0
9	345	-1	Interferência na Marginal Tietê sentido Castelo Branco pista local antes Ponte Nova Fepasa interdita uma faixa. #ZN	-1
10	346	1	SP 270 Raposo Tavares:\nTrecho São Paulo / Cotia:\nTráfego normal nos dois sentidos\nTempo instável e visibilidade red... https://t.co/PUyXT4t43	1
11	347	-1	RT @cge_sp: Estado de atenção para alagamentos nas Subprefeituras de Freguesia do Ó Casa Verde Santana Jaconã/Tremembé e Vila Maria/V	1
12	349	-1	Acidente com vítima na Ellis Maas próximo ao metrô capão bombeiros e PM no local.\n\n#mboimirim #sosmboimirim	-1
13	350	-1	Trânsito infernal causado por alagamento ?????????	1
14	351	-1	RT @radiotransitofm: ANCHIETA - Há obras bloqueando parcialmente ao trecho urbano da Anchieta para São Paulo. A interdição acontece na altu	-1

Figura 56 –Tweets por Tipo de Sentimento no Domínio do Transporte – Corpus Teste

Fonte: Autora

A Figura 57 apresenta as principais Hashtags dessa amostra.

Figura 57 – Principais Hashtags do Corpus Teste

Fonte: Autora

7.3. Considerações Finais do Capítulo

Neste estudo de caso foi implementado o método do Sistema de Análise de Sentimentos no Auxílio da Gestão das Cidades Inteligentes, no qual foram captadas informações provenientes da população de uma determinada região da cidade de São Paulo por meio da Rede Social Twitter.

O método desenvolvido para este trabalho de pesquisa utiliza o processo de mineração de dados com um filtro que classifica as informações por domínios de interesse da Cidade Inteligente. No caso do estudo de caso, foram utilizados os classificadores LinearSVC, Logistic Regression, Multinomial Naive Bayes e Random Forest, utilizando técnicas de extração de características BOW e TF-IDF, para classificar os domínios Transporte, Segurança e Educação.

A Tabela 5 apresenta um resumo do desempenho dos classificadores por técnica. O desempenho final dos classificadores foi avaliado pelas métricas acurácia e f1-score. No caso, o classificador que obteve o melhor resultado na classificação por Domínio foi o Logistic Regression usando a técnica TF-IDF, destacado em verde.

Tabela 5 – Tabela Resumo de Desempenho dos Classificadores por Domínio

	BOW		TF-IDF	
Domínio	Acurácia	F1 Score	Acurácia	F1 Score
LinearSVC	0.909091	0.90	0.878788	0.87
Logistic Regression	0.909091	0,90	0.909091	0.91
Naive Bayes	0.858586	0.85	0.888889	0.88
Random Forest	0.858586	0.84	0.848485	0.83

Fonte: Autora

Definido o classificador de Domínios de melhor desempenho, o Corpus de Teste foi validado, obtendo-se os seguintes resultados, conforme Tabela 6:

Tabela 6 – Tabela de Avaliação do Melhor Classificador por Domínio com o Corpus de Teste

CORPUS TESTE	TF-IDF	
Domínio	Acurácia	F1 Score
Logistic Regression	0.936170	0,93

Fonte: Autora

Também foi apresentado um estudo de Mineração de Dados para a Análise de Sentimentos no qual foram utilizados os mesmos classificadores da Mineração de Dados por Domínio para classificar os sentimentos em positivo, neutro e negativo. Os dados foram avaliados usando duas técnicas de extração de características de texto: Bag of Words e TF-IDF. Os resultados podem ser visualizados na Tabela 7.

Tabela 7 – Tabela Resumo de Desempenho dos Classificadores para Análise de Sentimentos

	BOW		TF-IDF	
Análise de Sentimentos	Acurácia	F1-Score	Acurácia	F1-Score
LinearSVC	0.857143	0.85	0.825397	0.80
Logistic Regression	0.841270	0.83	0.841270	0.83
Naive Bayes	0.809524	0.80	0.825397	0.82
Random Forest	0.746032	0.71	0.714286	0.67

Fonte: Autora

O desempenho final dos classificadores foi avaliado pelas métricas acurácia e f1-score. No caso, o classificador que obteve o melhor resultado na classificação para Análise de Sentimentos foi o LinearSVC, usando a técnica BOW, destacado em verde.

Definido o classificador para Análise de Sentimentos, de melhor desempenho, o Corpus de Teste foi validado, obtendo os seguintes resultados, conforme tabela 8:

Tabela 8 – Tabela de Avaliação do Melhor Classificador para Análise de Sentimentos
Usando o Corpus de Teste

CORPUS TESTE	BOW	
	Acúrcia	F1-Score
LinearSVC	0.840000	0.83

Fonte: Autora

Os resultados dos classificadores, em ambos os estudos, são satisfatórios, pois a capacidade humana de avaliação correta da subjetividade de um texto varia de 72% Wiebe et al. (2005) a 85% Golden (2011). Os experimentos demonstram que o objetivo foi atingido o que demonstra que o método escolhido é eficaz na classificação de tweets.

8. CONCLUSÕES E TRABALHOS FUTUROS

Neste capítulo são apresentadas as principais conclusões desta tese, bem como alguns trabalhos futuros que se mostraram úteis no decorrer do estudo.

8.1. Conclusões

Esta Tese contribuiu com o desenvolvimento de um novo método para um sistema capaz de captar, tratar, processar e avaliar os sentimentos contidos nas informações provenientes dos cidadãos de uma Cidade Inteligente. Esse Sistema atua como um sensor de informações, que pode trabalhar sozinho ou em conjunto com outros sensores físicos, tendo por objetivo auxiliar os gestores das Cidades Inteligentes nas tomadas de decisões específicas para cada domínio de interesse.

Analisando-se por outro foco, pode-se dizer que esta Tese possibilita que se insira a Análise de Sentimentos na gestão das chamadas Cidades Inteligentes, comportando-se como mais uma ferramenta de supervisão e controle dessas cidades. Considerando-se alguns dos benefícios advindos deste trabalho, é possível citar a identificação de ações, meios de prevenção e premeditação de possíveis adversidades, além da busca por melhorias na qualidade vida da população, auxiliando os gestores dessas cidades a tomar melhores decisões, considerando parte do conjunto de dados e informações disponíveis.

Tendo em vista que parcela significativa da população das Cidades Inteligentes já está conectada à Internet, torna-se possível realizar a captação de informações desses habitantes por meio das Redes Sociais. Sendo assim, o tratamento dos sentimentos das informações provenientes dos cidadãos das Cidades Inteligentes mostra-se um sensor a mais a ser explorado pelos gestores para auxílio nas tomadas de decisões.

Além disso, o método de Análise de Sentimentos possui um benefício de baixo custo de utilização, pois diferentemente dos outros sensores físicos, ele praticamente não apresenta custos de instalação e manutenção. Ele também é capaz de fornecer informações praticamente em tempo real, sendo possível, inclusive, a projeção e prevenção de eventos críticos.

Portanto, pode-se dizer que os objetivos propostos inicialmente foram atingidos, ou seja:

- Foi desenvolvido um Método de Desenvolvimento de um Sistema de Análise de Sentimentos para Auxílio na Gestão das Cidades Inteligentes (ASCI) para a análise das informações provenientes das Redes Sociais, incluindo-se um processo de pré-processamento dos dados, obtendo-se dados representativos e com qualidade; seguido por duas fases de mineração de dados, por meio de técnicas de aprendizado de máquina;
- Utilizaram-se dois processos de Mineração de Dados, sendo o primeiro para a filtragem das informações, considerando-se os domínios de interesse da Gestão das Cidades Inteligentes e o segundo processo de Mineração de Dados para a Análise de Sentimentos;
- Implementação de um estudo de caso, com a utilização de informações de cidadãos localizados em uma determinada região da cidade de São Paulo, provenientes da Rede Social Twitter, sendo esses dados captados e filtrados para o Domínio de Transportes, visando identificar situações críticas ou incomuns para o sistema de transporte de uma cidade.

8.2. Trabalhos Futuros

Durante a realização desta pesquisa, foram identificados alguns desafios e oportunidades que podem ser objetos de trabalhos futuros:

- Implementação de outros classificadores, por exemplo Redes Neurais, para novas avaliações de desempenho;
- Incremento do estudo de caso inserido-se ferramentas apropriadas para visualização dos resultados obtidos (gráficos, boxplots, mapas de palavras, gráfico de barras, etc.)
- Ampliar o número de domínios de interesse. Segundo (LIU, 2012), a Análise de Sentimentos é sensível em relação ao domínio utilizado no treinamento, no qual um conjunto de técnicas de extração bem como um classificador treinado para um determinado conjunto de dados não possui bom desempenho para outro domínio. Isso porque palavras ou expressões utilizadas para expressar uma opinião podem ser

diferentes para cada domínio e, portanto, cada etapa de extração de características e algoritmo de aprendizado apresenta resultados dependentes do domínio;

- Criação de um método automático de seleção de dados para os domínios de estudo;
- Implementar um sistema de predições de eventos futuros com base nos dados avaliados pela a Análise de Sentimentos;
- Utilizar a Análise de Sentimentos para que as informações avaliadas possam acionar automaticamente sistemas inteligentes. Por exemplo, semáforos que se auto-ajustam de acordo com o trânsito da região ou informações de acidentes, etc.

REFERÊNCIAS

- ABAKER, I.; HASHEM, T.; CHANG, V.; BADRUL, N.; ADEWOLE, K.; YAQOOB, I.; GANI, A.; AHMED, E.; CHIROMA, H. The role of big data in smart city. **International Journal of Information Management**, v. 36, n. 5, p. 748–758, 2016. Disponível em: <<http://dx.doi.org/10.1016/j.ijinfomgt.2016.05.002>>.
- AHMED, K. B.; BOUHORMA, M. Visual Sentiment Prediction with Transfer Learning and Big Data Analytics for smart cities. In: IEEE INTERNATIONAL COLLOQUIUM ON INFORMATION SCIENCE AND TECHNOLOGY (CiSt), 4. **Proceedings...**, section 2, p. 800–805, 2016.
- AHMED, K. B.; RADENSKI, A.; BOUHORMA, M.; BEN, M. Sentiment Analysis for Smart Cities: State of the Art and Opportunities. In: INT'L CONF. INTERNET COMPUTING AND INTERNET OF THINGS - ICOMP'16. **Proceedings...** p. 55–61, 2016.
- ALAVI, A. H.; JIAO, P.; BUTTLAR, W. G.; LAJNEF, N. Internet of Things-enabled smart cities: State-of-the-art and future trends. **Measurement**, v. 129, n. July, p. 589–606, 2018. Disponível em: <<https://doi.org/10.1016/j.measurement.2018.07.067>>.
- BBC. London Eye Olympic Twitter positivity lightshow launched. **BBC**, 2012. Disponível em: <<http://www.bbc.com/news/uk-england-london-18918318>>. Acesso em: 5 Mai. 2017.
- BECKER, K.; TUMITAN, D. **Introdução à mineração de opiniões: conceitos, aplicações e desafios**. Rio Grande do Sul: UFGRS, 2013.
- BERRONE, P.; RICART, J. E.; CARRASCO, C.; DUCH, A. CÁTEDRA SCHNEIDER ELECTRIC DE SOSTENIBILIDAD Y ESTRATEGIA. IESE Cities in Motion Index 2018. 2018. Disponível em: <<http://www.ieseinsight.com/fichaMaterial.aspx?pk=148539&idi=2&origen=3>>.
- BOCCHINI, B. **Apesar de melhorias, mobilidade em São Paulo demanda avanços**. **EBC Agência Brasil**. Disponível em: <<http://agenciabrasil.ebc.com.br/geral/noticia/2018-08/apesar-de-melhorias-mobilidade-em-sao-paulo-demanda-avancos>>. Acesso em: 26 Mar. 2019.
- BOYD, D.; ELLISON, N. **Social network sites: definition, history, and scholarship**. **IEEE Engineering Management Review**, v. 3, n. 38, p. 16-31, 2010.
- CAMILO, C. O.; SILVA, J. C.; **Mineração de dados: conceitos, tarefas, métodos e ferramentas**. Goiás: UFG, 2009.
- CARAGLIU, A.; BO, C. D.; NIJKAMP, P. Smart Cities in Europe Smart Cities in Europe. **Journal of Urban Technology**, v. 0732, n. August, p. 1–15, 2011.
- CET - SÃO PAULO - Companhia de Engenharia de Tráfego. **Visita Técnica**. 2018.

CHEN, F.; DENG, P.; WAN, J.; ZHANG, D.; VASILAKOS, A. V.; RONG, X. Data Mining for the Internet of Things: Literature Review and Challenges. **International Journal of Distributed Sensor Networks**, v. 2015, n. i, 2015. Disponível em: <<https://doi.org/10.1155/2015/431047>>

CORTÊS, S. C., PORCARO, R. M., LIFSCHITZ, S. **Mineração de Dados – Funcionalidades, Técnicas e Abordagens**. PUC-RioInf.MCC, Rio de Janeiro, MG, 2002

DAVIS, L. J.; OFFORD, K. P. Logistic regression. **Journal of Personality Assessment**, v. 68, n. 3, p. 497–507, 1997. Disponível em: <https://doi.org/10.1207/s15327752jpa6803_3>

DOMINGOS, P.; PAZZANI, M. On the Optimality of the Simple Bayesian Classifier under Zero-One Loss. **Machine Learning**, v. 29, p. 103–130, 1997.

FAYYAD, U.; PIATETSKY-SHAPIO, G.; SMYTH, P. From Data Mining to Knowledge Discovery in Data. **AI Magazine**, v. 17, n. 3, p. 37–54, 1996.

GANDOMI, A.; HAIDER, M. Beyond the hype : big data concepts , methods , and analytics. **International Journal of Information Management**, v. 35, n. 2, p. 137–144, 2015. Disponível em: <<http://dx.doi.org/10.1016/j.ijinfomgt.2014.10.007>>.

GAVILANES, M. F.; MARTÍNEZ, J. juncal; MÉNDEZ, S. G.; MONTENEGRO, E. C.; CASTAÑO, F. J. G. Differentiating users by language and location estimation in sentiment analysis of informal text during major public events. **Expert Systems with Applications**, v. 117, p. 15–28, 2019.

GIANG, N. K.; BLACKSTOCK, M.; LEA, R.; LEUNG, V. C. M. On Building Smart City IoT Applications: a Coordination-based Perspective. In: INTERNATIONAL WORKSHOP ON SMART CITIES: PEOPLE, TECHNOLOGY AND DATA, 2. **Anais...**, 2016.

HAJMOHAMMADI, M. S.; OTHMAN, Z. A. Opinion mining and sentiment analysis: a survey. **International Journal of Computers & Technology**, v. 2, n. 3, p. 2277–3061, 2012.

HAN, J.; KAMBER, M.; PEI, J. **Data mining concepts and techniques**. USA: Elsevier, 2012.

HAND, D.; MANNILA, H. ; SMYTH, P. **Principles of data mining**. USA: MIT, 2001

HOPFGARTNER, F. **Information systems**. USA: Springer, 2015

HRIDOY, S. A. A.; EKRAM, M. T.; ISLAM, M. S.; AHMED, F.; RAHMAN, R. M. Localized twitter opinion mining using sentiment analysis. **Decision Analytics**, v. 2, n. 1, p. 8, 2015. Disponível em: <<http://www.decisionanalyticsjournal.com/content/2/1/8>>.

IEEE Smart Cities. Disponível em: <<http://smartcities.ieee.org/about.html>>. Acesso em: 24 abr. 2017.

JAIN, U.; SANDHU, A. A review on the emotion detection from text using machine learning techniques. **International Journal of Current Engineering and Technology**, v. 5, n. 4, p. 2645–2650, 2015.

KEMP, Simon. Digital 2019. **Report**. HOOTSUITE, 2019. Disponível em: <<https://blog.hootsuite.com/simon-kemp-social-media/>>. Acesso em: 25 Mar. 2019.

LEWIS, L. **2018 Update: What Happens in an internet minute** by Lori Lewis Merge AllAccess.com. Disponível em: <<https://www.allaccess.com/merge/archive/28030/2018-update-what-happens-in-an-internet-minute>>. Acesso em: 10 fev. 2019.

LI, M.; CH, E.; CHONG, A.; SEE, S. The new eye of smart city : novel citizen sentiment analysis in twitter big data and visual analytics lab. In: INTERNATIONAL DOCTORAL INNOVATION CENTRE, UNIVERSITY OF NOTTINGHAM UNIVERSITY BUSINESS SCHOOL, CHINA NVIDIA ASIA PACIFIC AND JAPAN TECHNOLOGY CEN. INTERNATIONAL CONFERENCE ON AUDIO, LANGUAGE AND IMAGE PROCESSING (ICALIP), **Proceedings...**, p. 557–562, 2016.

LIU, B. Opinion mining and sentiment analysis: NLP meets social sciences. **STSC, Hawaii**, 2010. <http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html>

LIU, B. Sentiment analysis and opinion mining. **Morgan and Claypool Publishers**, n. May, 2012.

LUNARDI, A. D. C. **Classificação multiclasse de textos baseada em divisões binárias adaptadas ao domínio**. 2016. Dissertação (Mestrado) - Universidade Federal Fluminense, Niterói, 2016.

MANKAR, T.; HOTCHANDANI, T.; MADHWANI, M.; CHIDRAWAR, A.; LIFNA, C. S. Stock market prediction based on social sentiments using machine learning. In: INTERNATIONAL CONFERENCE ON SMART CITY AND EMERGING TECHNOLOGY, ICSCET 2018, **Proceedings...**, p. 1–3, 2018.

MOUSTAKA, V.; THEODOSIOU, Z.; VAKALI, A.; KOUNOUEDES, A.; ANTHOPOULOS, L. G.. **Enhancing social networking in smart cities: Privacy and security borderlines**. Technological Forecasting and Social Change. Elsevier, 2019 vol. 142(C), pages 285-300.

NIELSEN, F. A. A new: evolution of a word list for sentiment analysis in microblogs. In: ESWC2011 WORKSHOP ON 'MAKING SENSE OF MICROPOSTS': PROCEEDINGS OF THE ESWC2011 WORKSHOP ON 'MAKING SENSE OF MICROPOSTS': BIG THINGS COME IN SMALL PACKAGES, 2011. **Proceedings...** Nielsen, v. 718, p. 93–98, 2011.

ONU. World Urbanization Prospects The 2018 Revision. **Report**. Nova York, 2019. Disponível em: <<https://population.un.org/wup/Publications/Files/WUP2018-Report.pdf>>. Acesso em: 6 Jan. 2019.

ONUBR - NAÇÕES UNIDAS NO BRASIL. **Três milhões de pessoas por semana migram para centros urbanos, destaca agência da ONU Brasil**. Disponível em: <<https://nacoesunidas.org/tres-milhoes-de-pessoas-por-semana-migram-para-centros-urbanos-destaca-agencia-da-onu/>>. Acesso em: 21 nov. 2017.

OSMAN, A. M. S. A novel big data analytics framework for smart cities. **Future Generation Computer Systems**, v. 91, p. 620–633, 2019. Disponível em: <<https://doi.org/10.1016/j.future.2018.06.046>>.

PANG, B.; LEE, L. Opinion Mining and Sentiment Analysis. Foundations and Trends® in Information Retrieval, v. 1, n. 2, 91–231. Disponível em: <[doi:10.1561/15000000001n](https://doi.org/10.1561/15000000001n)>.

PANG, B.; LEE, L.; VAITHYANATHAN, S. Thumbs up?: sentiment classification using machine learning techniques. In: CONFERENCE ON EMPIRICAL METHODS IN NATURAL LANGUAGE PROCESSING, **Proceedings...** p. 79–86, 2002. Disponível em: <<http://portal.acm.org/citation.cfm?id=1118693.1118704>>.

PEREIRA, G. V.; BERNARDES, M. B.; BERNARDINI, F.; CAPPELLI, C.; GOMYDE, A. building a reference model and an evaluation method for cities of the brazilian network of smart and human cities. In: ANNUAL INTERNATIONAL CONFERENCE ON DIGITAL GOVERNMENT RESEARCH - DG.O'17, 18. **Proceedings...** p. 580–581, 2017. Disponível em: <<http://dl.acm.org/citation.cfm?doid=3085228.3085257>>.

PEREIRA, J. F. F. Social media text processing and semantic analysis for smart cities. 2017.

POZZI, F. A.; FERSINI, E. M. E.; LIU, B. **Sentiment analysis in social networks**. [s.l.]: Morgan & Claypool, 2017

PURI, M.; VARDE, A.; DU, X.; MELO, G. Smart governance through opinion mining of public reactions on ordinances. In: INTERNATIONAL CONFERENCE ON TOOLS WITH ARTIFICIAL INTELLIGENCE, ICTAI, **Proceedings...** v. 2018–November, p. 838–845, 2018.

RATHORE, M. M.; AHMAD, A.; PAUL, A. Big Data and Internet of Things: an asset for urban planning. In: INTERNATIONAL CONFERENCE ON BIG DATA APPLICATIONS AND SERVICES - BIGDAS '15. **Proceedings...** 2015, p. 58–65, 2015. Disponível em: <<http://dl.acm.org/citation.cfm?doid=2837060.2837067>>.

RATHORE, M. M.; AHMAD, A.; PAUL, A.; RHO, S. Urban planning and building smart cities based on the Internet of Things using Big Data analytics. **Computer Networks**, v. 101, p. 63–80, 2016. Disponível em: <<http://dx.doi.org/10.1016/j.comnet.2015.12.023>>.

REDE BRASILEIRA DE CIDADES INTELIGENTES & HUMANAS. Brasil 2030: cidades inteligentes e humanas. In: GOMYDE, A. et al. **Coletânea**. 2016.

RIBEIRO, F. BENEVENUTO, F. ARAÚJO, M. Métodos para análise de sentimentos em mídias sociais. 2013. Disponível em: <<http://homepages.dcc.ufmg.br/~fabricio/download/webmedia-short-course.pdf>>.

SAINANI, K. L. Logistic Regression. **Physical Medicine and Rehabilitation**, v. 6, p. 1157–1162, 2014. Disponível em: <<http://dx.doi.org/10.1016/j.pmrj.2014.10.006>>.

SANTANA, E. F. Z.; CHAVES, A. P.; GEROSA, M. A.; KON, F.; MILOJICIC, D. Software Platforms for Smart Cities: Concepts, Requirements, Challenges, and a Unified Reference Architecture. 2016. Disponível em: <<http://arxiv.org/abs/1609.08089>>.

STECANELLA, B. Donald Trump vs Hillary Clinton: sentiment analysis on Twitter mentions. **MonkeyLearn**, 2016. Disponível em: <https://monkeylearn.com/blog/trump-vs-hillary-sentiment-analysis-twitter-mentions>. Acesso em: 24 Fev. 2017.

SLUBAN, B.; SMAILOVIĆ, J.; BATTISTON, S.; MOZETIČ, I. Sentiment leaning of influential communities in social networks. **Computational Social Networks**, v. 2, n. 1, 2015.

SZERSZEN, D.; PALSSON, A. Sentiment classification in social media an analysis of methods and the impact of. 2016.

TABOADA, M.; BROOKE, J.; TOFILOSKI, M.; VOLL, K.; STEDE, M. Lexicon-based methods for sentiment analysis. **Computational Linguistics**, v. 37, n. 2, p. 267–307, 2011. Disponível em: <http://www.mitpressjournals.org/doi/abs/10.1162/COLI_a_00049>.

TOBERGTE, D. R.; CURTIS, S. Minicurso SBRC 2016. **Journal of Chemical Information and Modeling**, v. 53, n. 9, p. 1689–1699, 2013.

VIVIDPLANET. **Map Coordinates**. Disponível em: <<https://www.mapcoordinates.net/pt>>. Acesso em: 2 mar. 2019.

WITTEN, I. H.; EIBE, F. HALL, M. A. **Data mining practical machine learning tools and techniques**. [s.l]: 2012

XU, Z.; MEI, L.; CHOO, K.-K. R.; LV, Z.; HU, C.; LUO, X.; LIU, Y. Neurocomputing Mobile crowd sensing of human-like intelligence using social sensors: a survey. **Neurocomputing**, v. 279, p. 3–10, 2018. Disponível em: <<https://doi.org/10.1016/j.neucom.2017.01.127>>.

ZANELLA, A.; BUI, N.; CASTELLANI, A.; VANGELISTA, L.; ZORZI, M. Internet of Things for Smart Cities. **IEEE Internet of Things Journal**, v. 1, n. 1, p. 22–32, 2014. Disponível em: <<http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6740844>>.

ZHOU, G.; ZHAO, J.; ZENG, D. Sentiment Classification with Graph Co-Regularization. In: **International Conference on Computational Linguistics: Technical Papers**, p. 1331–1340, 2014. 25. **Proceedings...** Disponível em: <<http://www.aclweb.org/anthology/C14-1126>>.

Anexo I - Programa de captação de tweets através da API do Twitter

```
## TeseCaptacaoTwitter.ipynb ##
```

```
import oauth2 as oauth
import json
import time
```

```
# Consume:
CONSUMER_KEY      = '8mzFnkBT91D0SqfpxX52idNWW'
CONSUMER_SECRET   = 'zC35neJDV5jVEdDG2A6W2NLvtfTOyMowXF49EgmMT1ydawxxQA'
```

```
# Access:
ACCESS_TOKEN      = '1072868708643565570-yYsr3aDiWnKUu0k41VE9yPflp2lFCb'
ACCESS_SECRET     = 'ZdXvocfaiPdlxqrSvvgDnqJpFHBavPccDq1I5xVhUqapQ'
```

```
consumer = oauth.Consumer(key=CONSUMER_KEY, secret=CONSUMER_SECRET)
access_token = oauth.Token(key=ACCESS_TOKEN, secret=ACCESS_SECRET)
client = oauth.Client(consumer, access_token)
```

```
#Trânsito
```

```
#q='CETSP_%20OR%20CET%20OR%20trânsito%20OR%20acidente%20OR%20colisão%20OR%20engarramento%20OR%22batida%20de%20carro%22OR%20engavetamento%20OR%20atropelamento'
```

```
#q='capotamento%20OR%20semáforo%20OR%22farol%20quebrado%22OR%22trânsito%20bom%22OR%22trânsito%20livre%22OR%22tráfego%20normal%22OR%22trânsito%20normal%22'
```

```
#q='CETSP_%20OR%20trânsito%20OR%22trânsito%20bom%22OR%22trânsito%20livre%22OR%22tráfego%20normal%22OR%22trânsito%20normal%22OR%20Trânsito%20OR%22acidente%20com%20vítima%22OR%20motoqueiro%20OR%22interditada%20no%20sentido%22OR%20alagamento%20OR%20buraco'
```

```
# Segurança
```

```
#q='%22tráfico%20de%20drogas%22OR%20latrocínio%20OR%20assassinato%20OR%20pedofilia%20OR%20crime%20OR%20assalto%20OR%20roubo%20OR%20bullying%20OR%20PM%20OR%20polícia'
```

```
#q='estupro%20OR%20tiro%20OR%20tiroteio%20OR%20violência%20OR%20briga%20OR%20sequestro%20OR%20meliante%20OR%20socorro%20OR%20Tráfico'
```

```
#q='ronda%20OR%20arma%20OR%20matar%20OR%20morte%20OR%20revolver%20OR%20fuzilado%20OR%20atentado%20OR%20terrorista%20OR%22segurança%20pública%22'
```

```
#q='traficante%20OR%22tráfico%20de%20drogas%22OR%20tiro%20OR%20latrocínio%20OR%20assassinato%20OR%20pedofilia%20OR%20crime%20OR%20assalto%20OR%20roubo%20OR%20bullying'
```

```
#qf='PM%20OR%20políciaOR20segurança pública%20estupro%20OR%20policial%20OR%20tiro%20OR%20tiroteio%20OR%20violência%20OR%20briga%20OR%22torcida%20organizada%22OR%20sequestro'
```

```
#qf='meliante%20OR%20arma%20OR%20armado%20OR%20ronda%20OR%20assassinado%20OR%20fuzilado%20OR%20atentado%20OR%20terrorista'
```

```
# Educação
```

```
#q='Faculdade%20OR%20Universidade%20OR%22Diretor%20da%20escola%22OR%22D
iretora%20da%20escola%22OR%20Aulas%20OR%20Aula%20
#q='#EDUCAÇÃO%20OR%20educação%20OR%20MEC%20OR%20escola%20%OR%22Férias%2
0escolares%22OR%22Merenda%20escolar%22'
```

```
#url =
"https://api.twitter.com/1.1/search/tweets.json?q="+str(q)+"&count=100"
+"&lang=pt"
```

```
url =
"https://api.twitter.com/1.1/search/tweets.json?q="+str(q)+"&geocode=-
23.533773,-46.625290,10km"+"&lang=pt"+"&count=200"
response, data = client.request(url, "GET")
tweets = json.loads(data)
```

```
for tweet in tweets['statuses']:
    print (str(tweet))
```

```
{'created_at': 'Wed Apr 10 00:19:05 +0000 2019', 'id': 1115771145209380869, 'id_str':
'1115771145209380869', 'text': 'SP 125 Oswaldo Cruz:\nTráfego normal nos dois sentidos\nTempo
instável e visibilidade reduzida\nAtenção: Trecho de ser... https://t.co/oYFz07dGBp',
'truncated': True, 'entities': {'hashtags': [], 'symbols': [], 'user_mentions': [], 'urls':
[{'url': 'https://t.co/oYFz07dGBp', 'expanded_url':
'https://twitter.com/i/web/status/1115771145209380869', 'display_url':
'twitter.com/i/web/status/1...'}, {'indices': [117, 140]}]}, 'metadata': {'iso_language_code':
'pt', 'result_type': 'recent'}, 'source': '<a href="http://twitter.com" rel="nofollow">Twitter
Web Client</a>', 'in_reply_to_status_id': None, 'in_reply_to_status_id_str': None,
'in_reply_to_user_id': None, 'in_reply_to_user_id_str': None, 'in_reply_to_screen_name': None,
'user': {'id': 988909723, 'id_str': '988909723', 'name': 'DER-SP', 'screen_name': 'dersp',
'location': 'São Paulo, Brasil', 'description': 'Twitter Oficial do Departamento de Estradas
de Rodagem de São Paulo. Informações detalhadas ligue 0800 055 5510', 'url':
'https://t.co/FnBcUTpSzY', 'entities': {'url': {'urls': [{'url': 'https://t.co/FnBcUTpSzY',
'expanded_url': 'http://www.der.sp.gov.br', 'display_url': 'der.sp.gov.br', 'indices': [0,
23]}]}}, 'description': {'urls': []}}, 'protected': False, 'followers_count': 31460,
'friends_count': 13, 'listed_count': 205, 'created_at': 'Tue Dec 04 14:39:31 +0000 2012',
'favourites_count': 4, 'utc_offset': None, 'time_zone': None, 'geo_enabled': True, 'verified':
False, 'statuses_count': 138323, 'lang': 'pt', 'contributors_enabled': False, 'is_translator':
False, 'is_translation_enabled': False, 'profile_background_color': '003D4B',
'profile_background_image_url': 'http://abs.twimg.com/images/themes/theme1/bg.png',
'profile_background_image_url_https': 'https://abs.twimg.com/images/themes/theme1/bg.png',
'profile_background_tile': False, 'profile_image_url':
'http://pbs.twimg.com/profile_images/486213739301982210/vPRiG1U4_normal.jpeg',
'profile_image_url_https':
'https://pbs.twimg.com/profile_images/486213739301982210/vPRiG1U4_normal.jpeg',
'profile_link_color': '003D4B', 'profile_sidebar_border_color': 'FFFFFF',
'profile_sidebar_fill_color': 'DDEEF6', 'profile_text_color': '333333',
'profile_use_background_image': True, 'has_extended_profile': True, 'default_profile': False,
'default_profile_image': False, 'following': False, 'follow_request_sent': False,
'notifications': False, 'translator_type': 'none'}, 'geo': None, 'coordinates': None, 'place':
None, 'contributors': None, 'is_quote_status': False, 'retweet_count': 0, 'favorite_count': 1,
'favorited': False, 'retweeted': False, 'possibly_sensitive': False, 'lang': 'pt'}
```

Anexo II - Programas de implementação da Metodologia e Estudo de caso

a) Programa de Processamento e Mineração dos Dados referente ao Domínio dos temas de interesse

```
##Geral_Tese_Dominio.ipynb

import re # for regular expressions

import pandas as pd

pd.set_option("display.max_colwidth", 200)

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

import string

import nltk # for text manipulation

nltk.download('stopwords')

from wordcloud import WordCloud

from nltk.corpus import stopwords

from nltk.stem import WordNetLemmatizer

from nltk import tokenize

from sklearn.feature_extraction.text import TfidfVectorizer

%matplotlib inline

train = pd.read_csv('GeralTreinafinal.csv', encoding="ISO-8859-1")

test = pd.read_csv('GeralTestefinal.csv', encoding='ISO-8859-1')

# apresenta os 5 primeiros registros com o domínio 0 do arquivo de treinamento

train[train['dominio'] == 0].head(5)

train.head(5)

test.head(5)

# Número de dados e atributos

train.shape, test.shape

((329, 3), (141, 3))

# calcular das amostras por classe

train['dominio'] = train['dominio'].astype('category')
```



```

train["dominio"].value_counts()

# imprimir o gráfico representando as classes no treinamento
train.dominio.value_counts().plot(kind='bar')

# imprimir o gráfico representando as classes no teste
test.dominio.value_counts().plot(kind='bar')

# juntando os arquivos train e test para Limpeza dos Dados
combi = train.append(test, ignore_index=True)

def remove_pattern(input_txt, pattern):

    r = re.findall(pattern, input_txt)

    for i in r:

        input_txt = re.sub(i, ' ', str(input_txt))

    return input_txt

combi['clean_tweet'] = combi['tweet'].str.replace('@[\w]*', '', regex=True)

combi.head(15)

combi['clean_tweet'] = combi['clean_tweet'].str.replace("([[-|0-9])|(\n)|(...https://t.co/[\w]*)|([-|.|/?!,:;()\'#\]])", " ")

combi.head(15)

combi['clean_tweet'] = combi['clean_tweet'].apply(lambda x: ' '.join([w for w in x.split() if len(w)>2]))

combi.head(10)

tokenized_tweet = combi['clean_tweet'].apply(lambda x: x.split()) # tokenizing

tokenized_tweet.head(5)

from nltk.stem.porter import *

stemmer = PorterStemmer()

tokenized_tweet = tokenized_tweet.apply(lambda x: [stemmer.stem(i) for i in x]) # stemming

tokenized_tweet.head(5)

for i in range(len(tokenized_tweet)):

    tokenized_tweet[i] = ' '.join(tokenized_tweet[i])

```

```

combi['clean_tweet'] = tokenized_tweet

tokenized_tweet.head(5)

#nuvem de palavras

all_words = ' '.join([text for text in combi['clean_tweet']])

wordcloud = WordCloud(width=800, height=500, random_state=21,
max_font_size=110).generate(all_words)

plt.figure(figsize=(10, 7))

plt.imshow(wordcloud, interpolation="bilinear")

plt.axis('off')

plt.show()

normal_words = ' '.join([text for text in combi['clean_tweet'][combi['dominio'] ==
0]])

wordcloud = WordCloud(width=800, height=500, random_state=21,
max_font_size=110).generate(normal_words)

plt.figure(figsize=(10, 7))

plt.imshow(wordcloud, interpolation="bilinear")

plt.axis('off')

plt.show()

normal_words = ' '.join([text for text in combi['clean_tweet'][combi['dominio'] ==
1]])

wordcloud = WordCloud(width=800, height=500, random_state=21,
max_font_size=110).generate(normal_words)

plt.figure(figsize=(10, 7))

plt.imshow(wordcloud, interpolation="bilinear")

plt.axis('off')

plt.show()

normal_words = ' '.join([text for text in combi['clean_tweet'][combi['dominio'] ==
2]])

wordcloud = WordCloud(width=800, height=500, random_state=21,
max_font_size=110).generate(normal_words)

plt.figure(figsize=(10, 7))

plt.imshow(wordcloud, interpolation="bilinear")

plt.axis('off')

```

```

plt.show()

#Extracting Features from Cleaned Tweets

from sklearn.feature_extraction.text import TfidfVectorizer, CountVectorizer

import gensim

#Bag-of-Words Features

from nltk.corpus import stopwords

stopwords = set(stopwords.words('portuguese'))

bow_vectorizer = CountVectorizer(max_df=0.9, min_df=3, max_features=1000,
stop_words=stopwords,ngram_range=(1, 2))

bow = bow_vectorizer.fit_transform(combi['clean_tweet'])

bow_tot=bow

labels_tot=combi.dominio

bow_tot.shape

#TF-IDF Features

from nltk.corpus import stopwords

stopwords = set(stopwords.words('portuguese'))

tfidf_vectorizer = TfidfVectorizer(sublinear_tf=True,max_df=0.90, min_df=3,
norm='l2', max_features=1000, stop_words=stopwords,ngram_range=(1, 2))

tfidf = tfidf_vectorizer.fit_transform(combi['clean_tweet'])

tfidf_tot=tfidf

tfidf_tot.shape

# splitting data into training and validation set

from sklearn.model_selection import train_test_split

from sklearn.metrics import f1_score

train_bow = bow[:329,:]

test_bow = bow[329:,:]

# splitting data into training and validation set

## BOW ##

Xtrain_bow, Xvalid_bow, y_train, y_valid = train_test_split(train_bow,
train['dominio'],random_state=21, test_size=0.3)

## TFIDF ##

```

```

train_tfidf = tfidf[:329,:]

test_tfidf = tfidf[329,:]

Xtrain_tfidf = train_tfidf[y_train.index]

Xvalid_tfidf = train_tfidf[y_valid.index]

#### BOW - ACURÁCIA DOS DADOS 4 técnicas####

from sklearn.naive_bayes import MultinomialNB

from sklearn.linear_model import LogisticRegression

from sklearn.ensemble import RandomForestClassifier

from sklearn.svm import LinearSVC

from sklearn.model_selection import cross_val_score

models = [

    RandomForestClassifier(n_estimators=1000, max_depth=6, random_state=21),

    LinearSVC(),

    MultinomialNB(),

    LogisticRegression(solver='lbfgs', multi_class='multinomial', random_state=21),

]

CV = 5 # separação de aprendizado K-fold

cv_df = pd.DataFrame(index=range(CV * len(models)))

entries = []

for model in models:

    model_name = model.__class__.__name__

    accuracies = cross_val_score(model, bow_tot, labels_tot, scoring='accuracy',
cv=CV)

    for fold_idx, accuracy in enumerate(accuracies):

        entries.append((model_name, fold_idx, accuracy))

    cv_df = pd.DataFrame(entries, columns=['model_name', 'fold_idx', 'accuracy'])

import seaborn as sns

sns.boxplot(x='model_name', y='accuracy', data=cv_df)

sns.stripplot(x='model_name', y='accuracy', data=cv_df,

              size=8, jitter=True, edgecolor="gray", linewidth=2)

```

```

plt.show()

####TREINAMENTO Modelo BOW ####

####  MODELO LINEAR_SVC ###

from sklearn.model_selection import train_test_split

from sklearn import metrics

modelB1 = LinearSVC().fit(Xtrain_bow, y_train)

y_pred = modelB1.predict(Xvalid_bow)

dominio=[0, 1, 2]

print ('Acurácia: %f' %metrics.accuracy_score(y_valid, y_pred))

print(metrics.classification_report(y_valid, y_pred, dominio))

pd.crosstab(y_valid,y_pred)


####TREINAMENTO Modelo BOW ####

####  MODELO LOGISTIC_REGRESSION ###

from sklearn.linear_model import LogisticRegression

from sklearn import metrics

modelB2=LogisticRegression(solver='lbfgs',
multi_class='multinomial').fit(Xtrain_bow, y_train) # treinamento dos dados

y_pred = modelB2.predict(Xvalid_bow) # validação dos dados

print ('Acurácia: %f' %metrics.accuracy_score(y_valid, y_pred))

print(metrics.classification_report(y_valid, y_pred, dominio))

pd.crosstab(y_valid,y_pred)


####TREINAMENTO Modelo BOW ####

####  MODELO MULTINOMINAL NB ####

from sklearn.metrics import mean_squared_error

from sklearn.model_selection import train_test_split

from sklearn.feature_extraction.text import CountVectorizer

from sklearn.feature_extraction.text import TfidfTransformer

from sklearn.naive_bayes import MultinomialNB

```

```

modelB3 = MultinomialNB().fit(Xtrain_bow, y_train) # treinamento dos dados

y_pred = modelB3.predict(Xvalid_bow) # validação dos dados

print ('Acurácia: %f' %metrics.accuracy_score(y_valid, y_pred))

print(metrics.classification_report(y_valid, y_pred, dominio))

pd.crosstab(y_valid,y_pred)

####TREINAMENTO Modelo BOL ####

#### MODELO RANDOM FOREST CLASSIFIER ####

model = RandomForestClassifier(n_estimators=1000, max_depth=6, random_state=21)

modelB4= model.fit(Xtrain_bow, y_train) # treinamento dos dados

y_pred = modelB4.predict(Xvalid_bow) # validação dos dados

print ('Acurácia: %f' %metrics.accuracy_score(y_valid, y_pred))

print(metrics.classification_report(y_valid, y_pred, dominio))

pd.crosstab(y_valid,y_pred)

#### TFIDF - ACURÁCIA DOS DADOS 4 técnicas####

from sklearn.linear_model import LogisticRegression

from sklearn.ensemble import RandomForestClassifier

from sklearn.svm import LinearSVC

from sklearn.model_selection import cross_val_score

models = [

    RandomForestClassifier(n_estimators=1000, max_depth=6, random_state=21),

    LinearSVC(),

    MultinomialNB(),

    LogisticRegression(solver='lbfgs', multi_class='multinomial',random_state=21),

]

CV = 5 # separação de aprendizado K-fold

```

```

cv_df = pd.DataFrame(index=range(CV * len(models)))

entries = []

for model in models:

    model_name = model.__class__.__name__

    accuracies = cross_val_score(model, tfidf_tot, labels_tot, scoring='accuracy',
cv=CV)

    for fold_idx, accuracy in enumerate(accuracies):

        entries.append((model_name, fold_idx, accuracy))

    cv_df = pd.DataFrame(entries, columns=['model_name', 'fold_idx', 'accuracy'])

import seaborn as sns

sns.boxplot(x='model_name', y='accuracy', data=cv_df)

sns.stripplot(x='model_name', y='accuracy', data=cv_df,

              size=8, jitter=True, edgecolor="gray", linewidth=2)

plt.show()

####TREINAMENTO DOS Modelo TFIDF####

#### MODELO LOGISTIC_REGRESSION ###

from sklearn.linear_model import LogisticRegression

from sklearn import metrics

modelF2=LogisticRegression(solver='lbfgs',
multi_class='multinomial').fit(Xtrain_tfidf, y_train) # treinamento dos dados

y_pred = modelF2.predict(Xvalid_tfidf) # validação dos dados

```

```

print ('Acurácia: %f' %metrics.accuracy_score(y_valid, y_pred))

print(metrics.classification_report(y_valid, y_pred, dominio))

pd.crosstab(y_valid,y_pred)

####TREINAMENTO DOS Modelo TFIDF ####

#### MODELO MULTINOMINAL NB ####

from sklearn.naive_bayes import MultinomialNB

modelF3 = MultinomialNB().fit(Xtrain_tfidf, y_train) # treinamento dos dados

y_pred = modelF3.predict(Xvalid_tfidf) # validação dos dados

print ('Acurácia: %f' %metrics.accuracy_score(y_valid, y_pred))

print(metrics.classification_report(y_valid, y_pred, dominio))

pd.crosstab(y_valid,y_pred)

####TREINAMENTO DOS Modelo TFIDF ####

#### MODELO RANDOM FOREST CLASSIFIER ####

model = RandomForestClassifier(n_estimators=1000, max_depth=6, random_state=21)

modelF4= model.fit(Xtrain_tfidf, y_train) # treinamento dos dados

y_pred = modelF4.predict(Xvalid_tfidf) # validação dos dados

print ('Acurácia: %f' %metrics.accuracy_score(y_valid, y_pred))

print(metrics.classification_report(y_valid, y_pred, dominio))

pd.crosstab(y_valid,y_pred)

# O Modelo que apresentou melhor desempenho

y_pred= modelF2.predict(Xvalid_tfidf)

df = pd.DataFrame({'pred' : y_pred,

                   'dominio' : y_valid})

submission = df[['pred','dominio']]

submission.to_csv('dataF2.csv', index=False)

```



```

df['dominio'] = df['dominio'].astype('category')

df['dominio'].value_counts()

df['pred'] = df['pred'].astype('category')

df['pred'].value_counts()

# Gráfico comparando valores reais x valores preditos

import matplotlib.pyplot as pyplot

real = [49,23,27]

predito = [56,21,26]

# Definindo a largura das barras

barWidth = 0.25

# Aumentando o gráfico

plt.figure(figsize=(10,5))

# Definindo a posição das barras

r1 = np.arange(len(real))

r2 = [x + barWidth for x in r1]

# Criando as barras

plt.bar(r1, real, color='#808080', width=barWidth, label='Real')

plt.bar(r2, predito, color='#66CDAA', width=barWidth, label='Predito')

# Adiciando legendas as barras

plt.xlabel('Domínios')

plt.xticks([r + barWidth for r in range(len(real))], ['TRANSPORTE', 'SEGURANÇA',
'EDUCAÇÃO'] )

plt.ylabel('Nº Tweets')

plt.title('Gráfico Validação do Treinamento - Classificação dos Tweets por Domínios
(Real X Predito)')

# Criando a legenda e exibindo o gráfico

plt.legend()

plt.show()

```

```

# Pie chart, where the slices will be ordered and plotted counter-clockwise:

labels = 'Transporte', 'Segurança', 'Educação'

sizes = [49,23,27]

pyplot.axis('equal') # Equal aspect ratio ensures that pie is drawn as a circle.

pyplot.pie(sizes, labels=labels, autopct='%1.1f%%',

           shadow=True, startangle=90)

pyplot.title('Gráfico Treinamento % de Tweets por Domínio - Dados Reais')

pyplot.show()

# Pie chart, where the slices will be ordered and plotted counter-clockwise:

labels = 'Transporte', 'Segurança', 'Educação'

sizes = [56,21,26]

pyplot.axis('equal') # Equal aspect ratio ensures that pie is drawn as a circle.

pyplot.pie(sizes, labels=labels, autopct='%1.1f%%',

           shadow=True, startangle=90)

pyplot.title('Gráfico Treinamento % de Tweets por Domínio - Dados Preditos - Acurácia de 90.9%')
pyplot.show()

# Dados de teste aplicados no Modelo

y_test = modelF2.predict(test_tfidf)

test['pred'] = y_test
submission = test[['id','pred','dominio','tweet']]
submission.to_csv('modelF2.csv', index=False)

# Acurácia = (141-9)/141 =93.6%

y_validT=test['dominio']

print ('Acurácia: %f' %metrics.accuracy_score(y_validT, y_test))

print(metrics.classification_report(y_validT, y_test))

pd.crosstab(y_validT,y_test)

test['dominio'] = test['dominio'].astype('category')

test['dominio'].value_counts()

test['pred'] = test['pred'].astype('category')

test['pred'].value_counts()

```

```

# Gráfico comparando valores reais x valores preditos
import matplotlib.pyplot as pyplot
real = [59, 41, 41]
predito = [65, 33, 43]

# Definindo a largura das barras
barWidth = 0.25

# Aumentando o gráfico
plt.figure(figsize=(10,5))

# Definindo a posição das barras
r1 = np.arange(len(real))
r2 = [x + barWidth for x in r1]

# Criando as barras
plt.bar(r1, real, color='#808080', width=barWidth, label='Real')
plt.bar(r2, predito, color='#66CDAA', width=barWidth, label='Predito')

# Adiciando legendas as barras
plt.xlabel('Domínios')
plt.xticks([r + barWidth for r in range(len(real))], ['TRANSPORTE', 'SEGURANÇA', 'E
DUCAÇÃO'])
plt.ylabel('Nº Tweets')
plt.title('Gráfico Teste - Classificação dos Tweets por Domínios (Real X Predito)')

# Criando a legenda e exibindo o gráfico
plt.legend()
plt.show()

# Pie chart, where the slices will be ordered and plotted counter-clockwise:
labels = 'Transporte', 'Segurança', 'Educação'
sizes = [59, 41, 41]

pyplot.axis('equal') # Equal aspect ratio ensures that pie is drawn as a circle.

pyplot.pie(sizes, labels=labels, autopct='%1.1f%%',
            shadow=True, startangle=90)

pyplot.title('Gráfico Teste % de Tweets por Domínio - Dados Reais')

pyplot.show()

# Pie chart, where the slices will be ordered and plotted counter-clockwise:
labels = 'Transporte', 'Segurança', 'Educação'
sizes = [65, 33, 43]

pyplot.axis('equal') # Equal aspect ratio ensures that pie is drawn as a circle.

pyplot.pie(sizes, labels=labels, autopct='%1.1f%%', shadow=True, startangle=90)

pyplot.title('Gráfico Teste % de Tweets por Domínio - Dados Preditos - Acurácia de
93.6%')

pyplot.show()

test.head(20)

# extraindo Hashtags dos tweets de Teste referente ao domínio

# function to collect hashtags

def hashtag_extract(x):
    hashtags = []
    # Loop over the words in the tweet

```

```

for i in x:
    ht = re.findall(r"#(\w+)", i)

    hashtags.append(ht)

return hashtags

HT_dom = hashtag_extract(test['tweet'])

HT_dom = sum(HT_dom, [])

print(HT_dom)

a = nltk.FreqDist(HT_dom)
d = pd.DataFrame({'Hashtag': list(a.keys()),
                  'Count': list(a.values())})

# selecting top 20 most frequent hashtags
d = d.nlargest(columns="Count", n = 10)
plt.figure(figsize=(16,5))
ax = sns.barplot(data=d, x= "Hashtag", y = "Count")
ax.set(ylabel = 'Quantidade')
plt.show()

```

TFIDF - ACURÁCIA DOS DADOS 4 técnicas####

```

from sklearn.linear_model import LogisticRegression
from sklearn.ensemble import RandomForestClassifier
from sklearn.svm import LinearSVC
from sklearn.model_selection import cross_val_score

models = [

    RandomForestClassifier(n_estimators=1000, max_depth=6, random_state
=21),
    LinearSVC(),
    MultinomialNB(0.1),
    LogisticRegression(solver='lbfgs', multi_class='multinomial', random
_state=21),
]
CV = 5 # separação de aprendizado K-fold

cv_df = pd.DataFrame(index=range(CV * len(models)))
entries = []

for model in models:
    model_name = model.__class__.__name__
    accuracies = cross_val_score(model, tfidf_tot, labels_tot, scoring='a
ccuracy', cv=CV)

    for fold_idx, accuracy in enumerate(accuracies):
        entries.append((model_name, fold_idx, accuracy))

    cv_df = pd.DataFrame(entries, columns=['model_name', 'fold_idx', 'a
ccuracy'])

import seaborn as sns
sns.boxplot(x='model_name', y='accuracy', data=cv_df)
sns.stripplot(x='model_name', y='accuracy', data=cv_df,
              size=8, jitter=True, edgecolor="gray", linewidth=2)
plt.show()

```

```
cv_df.groupby('model_name').accuracy.mean()
```

b) Programa de Processamento e Mineração dos Dados referente à Análise de Sentimentos no Domínio do Transporte

```
# Análise de Sentimwntos
import re # for regular expressions
import pandas as pd
pd.set_option("display.max_colwidth", 200)
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import string
import nltk # for text manipulation
nltk.download('stopwords')

from wordcloud import WordCloud
from nltk.corpus import stopwords
from nltk.stem import WordNetLemmatizer
from nltk import tokenize
from sklearn.feature_extraction.text import TfidfVectorizer

%matplotlib inline

train = pd.read_csv('GeralTranspTreinal.csv', encoding="ISO-8859-1")
test = pd.read_csv('GeralTranspTeste.csv', encoding='ISO-8859-1')

# apresenta os 5primeiros registros com o label do arquivo de treinamen
to
train[train['label'] == 0].head(5)

# Número de dados e atributos
train.shape, test.shape

train['label'] = train['label'].astype('category')

train["label"].value_counts()

train.label.value_counts().plot(kind='bar')

test.label.value_counts().plot(kind='bar')

#Data Cleaning
combi = train.append(test, ignore_index=True)

submission = combi[['id','label','Tweet']]
submission.to_csv('senti09042019.csv', index=False)
```

```

combi.head(5)
combi.shape

def remove_pattern(input_txt, pattern):
    r = re.findall(pattern, input_txt)
    for i in r:
        input_txt = re.sub(i, ' ', str(input_txt))

    return input_txt

combi['clean_tweet'] = combi['Tweet'].str.replace('@[\w]*', '', regex=True)
combi.head(15)

combi['clean_tweet'] = combi['clean_tweet'].str.replace("((-|0-9))|(\n)|(...https://t.co/[\w]*)|([-|./?!,:;()\'#\])", " ")

combi.head(15)

combi['clean_tweet'] = combi['clean_tweet'].apply(lambda x: ' '.join([w
    for w in x.split() if len(w)>2]))
combi.head(10)

tokenized_tweet = combi['clean_tweet'].apply(lambda x: x.split()) # tok
enizing

tokenized_tweet.head(5)

from nltk.stem.porter import *
stemmer = PorterStemmer()

tokenized_tweet = tokenized_tweet.apply(lambda x: [stemmer.stem(i) for
i in x]) # stemming
tokenized_tweet.head(5)

for i in range(len(tokenized_tweet)):
    tokenized_tweet[i] = ' '.join(tokenized_tweet[i])

combi['clean_tweet'] = tokenized_tweet

tokenized_tweet.head(5)

all_words = ' '.join([text for text in combi['clean_tweet']])

#import matplotlib.pyplot as plt

wordcloud = WordCloud(width=800, height=500, random_state=21, max_font_
size=110).generate(all_words)

plt.figure(figsize=(10, 7))
plt.imshow(wordcloud, interpolation="bilinear")
plt.axis('off')
plt.show()

normal_words = ' '.join([text for text in combi['clean_tweet'][combi['la
bel'] == 0]])

```

```
wordcloud = WordCloud(width=800, height=500, random_state=21, max_font_
size=110).generate(normal_words)
plt.figure(figsize=(10, 7))
plt.imshow(wordcloud, interpolation="bilinear")
plt.axis('off')
plt.show()
```

```
normal_words = ' '.join([text for text in combi['clean_tweet'][combi['la
bel'] == 1]])
```

```
wordcloud = WordCloud(width=800, height=500, random_state=21, max_font_
size=110).generate(normal_words)
plt.figure(figsize=(10, 7))
plt.imshow(wordcloud, interpolation="bilinear")
plt.axis('off')
plt.show()
```

```
normal_words = ' '.join([text for text in combi['clean_tweet'][combi['la
bel'] == -1]])
```

```
wordcloud = WordCloud(width=800, height=500, random_state=21, max_font_
size=110).generate(normal_words)
plt.figure(figsize=(10, 7))
plt.imshow(wordcloud, interpolation="bilinear")
plt.axis('off')
plt.show()
```

```
#Extracting Features from Cleaned Tweets
```

```
from sklearn.feature_extraction.text import TfidfVectorizer, CountVecto
rizer
import gensim
```

```
#Bag-of-Words Features
```

```
from nltk.corpus import stopwords
stopwords = set(stopwords.words('portuguese'))
```

```
bow_vectorizer = CountVectorizer(max_df=0.9, min_df=3, max_features=100
0, stop_words=stopwords, ngram_range=(1, 2))
bow = bow_vectorizer.fit_transform(combi['clean_tweet'])
bow_tot=bow
```

```
labels_tot=combi.label
```

```
bow_tot.shape
```

```
#TF-IDF Features
```

```
from nltk.corpus import stopwords
stopwords = set(stopwords.words('portuguese'))
```

```
tfidf_vectorizer = TfidfVectorizer(sublinear_tf=True, max_df=0.90, min_
df=3, norm='l2', max_features=1000, stop_words=stopwords, ngram_range=(1
, 2))
tfidf = tfidf_vectorizer.fit_transform(combi['clean_tweet'])
```

```
tfidf_tot=tfidf
tfidf_tot.shape
```

```
from sklearn.model_selection import train_test_split
from sklearn.metrics import f1_score
```

```
train_bow = bow[:210,:]
test_bow = bow[210:::]
```

```
# splitting data into training and validation set
```

```
## BOW ##
```

```
Xtrain_bow, Xvalid_bow, y_train, y_valid = train_test_split(train_bow,
train['label'],random_state=21, test_size=0.3)
```

```
## TFIDF ##
```

```
train_tfidf = tfidf[:210,:]
test_tfidf = tfidf[210:::]
```

```
Xtrain_tfidf = train_tfidf[y_train.index]
Xvalid_tfidf = train_tfidf[y_valid.index]
```

```
#### BOW - ACURÁCIA DOS DADOS 4 técnicas####
```

```
from sklearn.naive_bayes import MultinomialNB
from sklearn.linear_model import LogisticRegression
from sklearn.ensemble import RandomForestClassifier
from sklearn.svm import LinearSVC
from sklearn.model_selection import cross_val_score
```

```
models = [
```

```
    RandomForestClassifier(n_estimators=1000, max_depth=6, random_state
=21),
    LinearSVC(),
    MultinomialNB(0.1),
    LogisticRegression(solver='lbfgs', multi_class='multinomial',random
_state=21),
]
```

```
CV = 5 # separação de aprendizado K-fold
```

```
cv_df = pd.DataFrame(index=range(CV * len(models)))
entries = []
```

```
for model in models:
```

```
    model_name = model.__class__.__name__
    accuracies = cross_val_score(model, bow_tot, labels_tot, scoring='acc
uracy', cv=CV)
```

```
    for fold_idx, accuracy in enumerate(accuracies):
        entries.append((model_name, fold_idx, accuracy))
```

```
    cv_df = pd.DataFrame(entries, columns=['model_name', 'fold_idx', 'a
ccuracy'])
```

```
import seaborn as sns
```



```
sns.boxplot(x='model_name', y='accuracy', data=cv_df)
sns.stripplot(x='model_name', y='accuracy', data=cv_df,
              size=8, jitter=True, edgecolor="gray", linewidth=2)
plt.show()
```

```
####TREINAMENTO Modelo BOW####
```

```
#### MODELO LINEAR_SVC ###
```

```
from sklearn.model_selection import train_test_split
from sklearn import metrics
```

```
modelB1 = LinearSVC().fit(Xtrain_bow, y_train)
```

```
y_pred = modelB1.predict(Xvalid_bow)
```

```
print ('Acurácia: %f' %metrics.accuracy_score(y_valid, y_pred))
print(metrics.classification_report(y_valid, y_pred))
pd.crosstab(y_valid,y_pred)
```

```
####TREINAMENTO Modelo BOW####
```

```
#### MODELO LOGISTIC_REGRESSION ###
```

```
from sklearn.linear_model import LogisticRegression
from sklearn import metrics
```

```
modelB2=LogisticRegression(solver='lbfgs', multi_class='multinomial').f
it(Xtrain_bow, y_train) # treinamento dos dados
```

```
y_pred = modelB2.predict(Xvalid_bow) # validação dos dados
```

```
print ('Acurácia: %f' %metrics.accuracy_score(y_valid, y_pred))
print(metrics.classification_report(y_valid, y_pred))
pd.crosstab(y_valid,y_pred)
```

```
####TREINAMENTO Modelo BOW####
```

```
#### MODELO MULTINOMINAL NB ####
```

```
from sklearn.metrics import mean_squared_error
```

```
from sklearn.model_selection import train_test_split
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.feature_extraction.text import TfidfTransformer
from sklearn.naive_bayes import MultinomialNB
```

```

modelB3 = MultinomialNB().fit(Xtrain_bow, y_train) # treinamento dos da
dos

y_pred = modelB3.predict(Xvalid_bow) # validação dos dados

print ('Acurácia: %f' %metrics.accuracy_score(y_valid, y_pred))

print(metrics.classification_report(y_valid, y_pred))

pd.crosstab(y_valid,y_pred)

####TREINAMENTO Modelo BOW ####

#### MODELO RANDOM FOREST CLASSIFIER ####

model = RandomForestClassifier(n_estimators=1000, max_depth=6, random_s
tate=21)

modelB4= model.fit(Xtrain_bow, y_train) # treinamento dos dados

y_pred = modelB4.predict(Xvalid_bow) # validação dos dados

print ('Acurácia: %f' %metrics.accuracy_score(y_valid, y_pred))

print(metrics.classification_report(y_valid, y_pred))

pd.crosstab(y_valid,y_pred)

#### TFIDF - ACURÁCIA DOS DADOS 4 técnicas####

from sklearn.linear_model import LogisticRegression
from sklearn.ensemble import RandomForestClassifier
from sklearn.svm import LinearSVC
from sklearn.model_selection import cross_val_score

models = [

    RandomForestClassifier(n_estimators=1000, max_depth=6, random_state
=21),
    LinearSVC(),
    MultinomialNB(0.1),
    LogisticRegression(solver='lbfgs', multi_class='multinomial',random
_state=21),
]
CV = 5 # separação de aprendizado K-fold

cv_df = pd.DataFrame(index=range(CV * len(models)))
entries = []

for model in models:
    model_name = model.__class__.__name__
    accuracies = cross_val_score(model, tfidf_tot, labels_tot, scoring='a
ccuracy', cv=CV)

```

```

    for fold_idx, accuracy in enumerate(accuracies):
        entries.append((model_name, fold_idx, accuracy))

    cv_df = pd.DataFrame(entries, columns=['model_name', 'fold_idx', 'accuracy'])

import seaborn as sns
sns.boxplot(x='model_name', y='accuracy', data=cv_df)
sns.stripplot(x='model_name', y='accuracy', data=cv_df,
              size=8, jitter=True, edgecolor="gray", linewidth=2)
plt.show()

####TREINAMENTO DOS MODELOS TFIDF (Term Frequency times inverse document frequency.)####

#### MODELO LINEAR_SVC ###

from sklearn.model_selection import train_test_split
from sklearn import metrics

modelF1 = LinearSVC().fit(Xtrain_tfidf, y_train)

y_pred = modelF1.predict(Xvalid_tfidf) # validação dos dados

print ('Acurácia: %f' %metrics.accuracy_score(y_valid, y_pred))
print(metrics.classification_report(y_valid, y_pred))
pd.crosstab(y_valid,y_pred)

####TREINAMENTO DOS Modelo TFIDF ####

#### MODELO LOGISTIC_REGRESSION ###

from sklearn.linear_model import LogisticRegression
from sklearn import metrics

modelF2=LogisticRegression(solver='lbfgs', multi_class='multinomial').fit(Xtrain_tfidf, y_train) # treinamento dos dados

y_pred = modelF2.predict(Xvalid_tfidf) # validação dos dados

print ('Acurácia: %f' %metrics.accuracy_score(y_valid, y_pred))
print(metrics.classification_report(y_valid, y_pred))
pd.crosstab(y_valid,y_pred)

####TREINAMENTO DOS Modelo TFIDF ####

#### MODELO MULTINOMINAL NB ####

from sklearn.naive_bayes import MultinomialNB

modelF3 = MultinomialNB(0.1).fit(Xtrain_tfidf, y_train) # treinamento dos dados

```

```

y_pred = modelF3.predict(Xvalid_tfidf) # validação dos dados

print ('Acurácia: %f' %metrics.accuracy_score(y_valid, y_pred))
print(metrics.classification_report(y_valid, y_pred))
pd.crosstab(y_valid,y_pred)

####TREINAMENTO DOS Modelo TFIDF ####

#### MODELO RANDOM FOREST CLASSIFIER ####

model = RandomForestClassifier(n_estimators=1000, max_depth=6, random_s
tate=21)

modelF4= model.fit(Xtrain_tfidf, y_train) # treinamento dos dados
y_pred = modelF4.predict(Xvalid_tfidf) # validação dos dados

print ('Acurácia: %f' %metrics.accuracy_score(y_valid, y_pred))
print(metrics.classification_report(y_valid, y_pred))
pd.crosstab(y_valid,y_pred)

# O Modelo que apresentou melhor desempenho ModelB1
y_pred= modelB1.predict(Xvalid_bow)

df = pd.DataFrame({'pred' : y_pred,
                   'label' : y_valid})

submission = df[['pred','label']]
submission.to_csv('dataB1.csv', index=False)

df['label'] = df['label'].astype('category')
df['label'].value_counts()

df['pred'] = df['pred'].astype('category')
df['pred'].value_counts()

# TREIN Gráfico comparando valores reais x valores preditos
import matplotlib.pyplot as pyplot
real = [33,19,11]
predito = [34,18,11]

# Definindo a largura das barras
barWidth = 0.25

```

```

# Aumentando o gráfico
plt.figure(figsize=(10,5))

# Definindo a posição das barras
r1 = np.arange(len(real))
r2 = [x + barWidth for x in r1]

# Criando as barras
plt.bar(r1, real, color='#808080', width=barWidth, label='Real')
plt.bar(r2, predito, color='#66CDAA', width=barWidth, label='Predito')

# Adiciando legendas as barras
plt.xlabel('Transporte')
plt.xticks([r + barWidth for r in range(len(real))], ['Negativo', 'Neutro', 'Positivo'])
plt.ylabel('Nº Tweets')
plt.title('Gráfico Validação Treinamento - Classificação dos Tweets por Sentimentos (Real X Predito)')

# Criando a legenda e exibindo o gráfico
plt.legend()
plt.show()

```

```

import matplotlib.pyplot as pyplot

# Pie chart, where the slices will be ordered and plotted counter-clock wise:
labels = 'Negativo', 'Positivo', 'Neutro'
sizes = [33,19,11]
pyplot.axis('equal') # Equal aspect ratio ensures that pie is drawn as a circle.

pyplot.pie(sizes, labels=labels, autopct='%1.1f%%',
            shadow=True, colors=['red', 'blue', 'yellow'], startangle=90)

pyplot.title('Gráfico Treinamento % de Tweets por Domínio - Dados Reais')

pyplot.show()

```

```

import matplotlib.pyplot as pyplot

# Pie chart, where the slices will be ordered and plotted counter-clock wise:
labels = 'Negativo', 'Positivo', 'Neutro'
sizes = [34,18,11]
pyplot.axis('equal') # Equal aspect ratio ensures that pie is drawn as a circle.

pyplot.pie(sizes, labels=labels, autopct='%1.1f%%',
            shadow=True, colors=['red', 'blue', 'yellow'], startangle=90)

pyplot.title('Gráfico Treinamento % de Tweets por Domínio - Dados Preditos - Acurácia de 85,7%')

pyplot.show()

```

```

# O Modelo que apresentou melhor desempenho
y_test = modelB1.predict(test_bow)

test['pred'] = y_test
submission = test[['id','pred','label','Tweet']]
submission.to_csv('modelB3111.csv', index=False)

y_validT=test['label']

print ('Acurácia: %f' %metrics.accuracy_score(y_validT, y_test))

print(metrics.classification_report(y_validT, y_test))

pd.crosstab(y_validT,y_test)

test['label'] = test['label'].astype('category')

test['label'].value_counts()

test['pred'] = test['pred'].astype('category')

test['pred'].value_counts()

# Gráfico comparando valores reais x valores preditos
import matplotlib.pyplot as pyplot
real = [21, 13, 16]
predito = [22,11,17]

# Definindo a largura das barras
barWidth = 0.25

# Aumentando o gráfico
plt.figure(figsize=(10,5))

# Definindo a posição das barras
r1 = np.arange(len(real))
r2 = [x + barWidth for x in r1]

# Criando as barras
plt.bar(r1, real, color='#808080', width=barWidth, label='Real')
plt.bar(r2, predito, color='#66CDAA', width=barWidth, label='Predito')

# Adiciando legendas as barras
plt.xlabel('Transporte')
plt.xticks([r + barWidth for r in range(len(real))], ['Negativo', 'Neutro', 'Positivo'] )
plt.ylabel('Nº Tweets')
plt.title('Gráfico Teste - Classificação dos Tweets por Sentimentos (Real X Predito)')

# Criando a legenda e exibindo o gráfico
plt.legend()
plt.show()

import matplotlib.pyplot as pyplot

```

```
# Pie chart, where the slices will be ordered and plotted counter-clock
wise:
labels = 'Negativo', 'Positivo', 'Neutro'
sizes = [21, 16, 13]
pyplot.axis('equal') # Equal aspect ratio ensures that pie is drawn as
a circle.

pyplot.pie(sizes, labels=labels, autopct='%1.1f%%',
           shadow=True, colors=['red', 'blue', 'yellow'], startangle=90)

pyplot.title('Gráfico Teste % de Tweets por Domínio - Dados Reais')

pyplot.show()
```

```
import matplotlib.pyplot as pyplot
```

```
# Pie chart, where the slices will be ordered and plotted counter-clock
wise:
labels = 'Negativo', 'Positivo', 'Neutro'
sizes = [22, 17, 11]
pyplot.axis('equal') # Equal aspect ratio ensures that pie is drawn as
a circle.

pyplot.pie(sizes, labels=labels, autopct='%1.1f%%',
           shadow=True, colors=['red', 'blue', 'yellow'], startangle=90)

pyplot.title('Gráfico Teste % de Tweets por Domínio - Dados Preditos -
Acurácia de 84%')

pyplot.show()
```

```
test.head(15)
```

```
# extraindo Hashtags dos tweets de Teste referente ao label (SENTIMENTO
)
```

```
# function to collect hashtags
```

```
def hashtag_extract(x):
    hashtags = []
    # Loop over the words in the tweet
    for i in x:
        ht = re.findall(r"#(\w+)", i)

        hashtags.append(ht)

    return hashtags

HT_sent = hashtag_extract(test['Tweet'])

HT_sent = sum(HT_sent, [])

print(HT_sent)
```

```
a = nltk.FreqDist(HT_sent)
d = pd.DataFrame({'Hashtag': list(a.keys()),
```

```
        'Count': list(a.values()))})

# selecting top 20 most frequent hashtags
d = d.nlargest(columns="Count", n = 10)
plt.figure(figsize=(16,5))
ax = sns.barplot(data=d, x= "Hashtag", y = "Count")
ax.set(ylabel = 'Count')
plt.show()
```