UNIVERSIDADE FEDERAL DE SANTA CATARINA
CENTRO DE CIÊNCIAS, TECNOLOGIAS E SAÚDE

Luciana Regina Bencke

CLASSIFICAÇÃO AUTOMÁTICA DE MENSAGENS DE REDES SOCIAIS EM DIMENSÕES DOS MODELOS DE CIDADES INTELIGENTES

Araranguá

2019

Luciana Regina Bencke

CLASSIFICAÇÃO AUTOMÁTICA DE MENSAGENS DE REDES SOCIAIS EM DIMENSÕES DOS MODELOS DE CIDADES INTELIGENTES

Dissertação submetida ao PPGTIC - Programa de Pós-Graduação em Tecnologias da Informação e Comunicação para a obtenção do Grau de mestre.
Orientador: Prof. Cristian Cechinel, Dr.

Araranguá

2019

Ficha de identificação da obra elaborada pelo autor,
através do Programa de Geração Automática da Biblioteca Universitária da UFSC.

Bencke, Luciana Regina
Classificação automática de mensagens de redes sociais em dimensões dos modelos de Cidades Inteligentes / Luciana Regina Bencke ; orientador, Cristian Cechinel, 2019.
134 p.

Dissertação (mestrado) - Universidade Federal de Santa Catarina, Campus Araranguá, Programa de Pós Graduação em Tecnologias da Informação e Comunicação, Araranguá, 2019.

Inclui referências.

1. Tecnologias da Informação e Comunicação. 2. Aprendizagem de Máquina. 3. Redes Sociais. 4. Cidades Inteligentes. 5. ISO 37120. I. Cechinel, Cristian. II. Universidade Federal de Santa Catarina. Programa de Pós-Graduação em Tecnologias da Informação e Comunicação. III. Título.

Luciana Regina Bencke

CLASSIFICAÇÃO AUTOMÁTICA DE MENSAGENS DE REDES SOCIAIS EM DIMENSÕES DOS MODELOS DE CIDADES INTELIGENTES

Esta Dissertação foi julgada adequada para obtenção do Título de "Mestre em Tecnologias da Informação e Comunicação", e aprovada em sua forma final pelo PPGTIC - Programa de Pós-Graduação em Tecnologias da Informação e Comunicação

Araranguá, 01 de abril de 2019.

Prof.ª Andréa Cristina Trierweiller, Dr.ª
Coordenador do Curso

Banca Examinadora:

Prof. Cristian Cechinel, Dr.
Orientador
Universidade Federal de Santa Catarina

Prof.ª Merisandra Côrtes de Mattos Garcia, Dr.ª
Universidade do Extremo Sul Catarinense

Prof.ª Silvia Modesto Nassar, Dr.ª
Universidade Federal de Santa Catarina
(videoconferência)

Prof. Alexandre Leopoldo Gonçalves, Dr.
Universidade Federal de Santa Catarina

Este trabalho é dedicado aos meus pais, Nougathe e Elemar, e à minha querida Lordi (*in memoriam*).

AGRADECIMENTOS

Agradeço à família, amigos, colegas, professores e a todos que de alguma forma me ajudaram.

Aos meus pais, Nougathe e Elemar, pela educação que me foi dada e pelas palavras de apoio e compreensão desde o início deste trabalho.

À minha madrinha e mãe do coração, Lordi, que mesmo não estando mais aqui, foi fundamental na minha educação e me inspira todos os dias a ser uma pessoa melhor.

Ao meu companheiro Osvaldo, pelo seu apoio nesta nova jornada no ramo da pesquisa e educação, e por acreditar tanto em mim e na minha capacidade de superação das dificuldades.

À minha irmã Cristina por estar sempre presente, em especial auxiliando meus pais quando não pude estar por perto devido à distância e dedicação ao mestrado.

Ao meu amigo Luis Henrique, por suas palavras sempre motivadoras e realistas.

Ao meu orientador, Cristian, por ter me aceito como sua orientanda, por todo seu apoio, dedicação e pelos ensinamentos que para sempre servirão de base e exemplo para o desenvolvimento de uma almejada trajetória na pesquisa e ensino. Ao professor Roberto Munoz pelo auxílio com as imagens e apoio no artigo final.

À Fundação de Amparo à Pesquisa e Inovação do Estado de Santa Catarina (FAPESC) e à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pela bolsa de mestrado que me foi concedida.

À Universidade Federal de Santa Catarina por propiciar este passo tão importante em minha vida e por me preparar para contribuir no desenvolvimento da educação e pesquisa no Brasil, pontos que podem efetivamente transformar a sociedade.

Finalmente, agradeço ao contínuo fluxo da vida, que nos apresenta as dificuldades para que as encaremos como transformadoras, pois são, na verdade, formas de aprimorar nossa técnica, força e criatividade.

Obrigada!

"All that is necessary for faith is the belief that by doing our best we shall come nearer to success and that success in our aims (the improvement of the lot of mankind, present and future) is worth attaining."

(Rosalind Franklin, 1940)

"I have climbed the highest mountains
I have run through the fields
Only to be with you...
I have run, I have crawled
I have scaled these city walls
These CITY WALLS
Only to be with you
But I still haven't found What I'm looking for... "

(U2, 1987)

RESUMO

Uma cidade inteligente pode ser definida como uma cidade de alta tecnologia com vários recursos para resolver ou mitigar problemas normalmente gerados pela rápida urbanização. Diferentes modelos de indicadores foram desenvolvidos para acompanhar a evolução das cidades na busca por tornarem-se Cidades Inteligentes. Um exemplo é o padrão 37120 da Organização Internacional para Padronização (ISO), que propõe um conjunto de dimensões e indicadores para serviços e qualidade de vida para cidades e comunidades sustentáveis. Tem sido comum encontrar nas redes sociais perfis oficiais de organizações e entidades governamentais relacionadas aos serviços que elas fornecem ou pelos quais são responsáveis (água, resíduos, transporte, eventos culturais, etc.). Os cidadãos interagem com estes perfis diretamente para comunicar problemas sobre os serviços da cidade. O presente trabalho objetiva aplicar algoritmos de aprendizado de máquina sobre os dados urbanos gerados pelas redes sociais, a fim de criar classificadores para categorizar automaticamente as mensagens dos cidadãos de acordo com as diferentes dimensões dos serviços das cidades. Para tanto, dois conjuntos distintos de textos em português foram coletados de duas redes sociais: Twitter (1.950 tweets) e Colab (65.066 postagens). Os textos foram mapeados de acordo com as diferentes categorias ISO 37120, pré-processados e minerados por meio de 11 algoritmos implementados na Scikit-Learn. Os primeiros resultados indicaram a viabilidade da proposta, com os modelos alcançando médias em torno de 59% para a F1-macro e 75% para a F1-micro ao usar *Linear Support Vector Classification* (LSVC) e *Complement Naive Bayes* (CNB). No entanto, como os conjuntos de dados estavam altamente desbalanceados, os desempenhos dos modelos variam significativamente para cada categoria ISO, com os melhores resultados de F1-score ocorrendo para *Transporte* (87%), *Energia* (83%) e *Águas Residuais* (74%). Os classificadores gerados neste trabalho podem ser integrados à diversos serviços e sistemas da cidade, tais como: sistemas de suporte à decisão governamental, sistemas de reclamações para cidadãos, painéis comunitários, centrais de polícia, empresas de transporte, produtores culturais, agências ambientais e empresas de reciclagem.

Palavras-chave: Classificação de tópicos. Aprendizagem de Máquina. Redes Sociais. Serviços para Cidades Inteligentes. ISO 37120.

ABSTRACT

A Smart City can be defined as a high-tech city with several capabilities to strategically solve (or mitigate) problems normally generated by rapid urbanization. Different models of indicators have been developed to follow cities' development to become a Smart City. An example of such model is the standard 37120 from the International Organization for Standardization (ISO) that proposes a set of dimensions and indicators for services and quality of life for sustainable cities and communities. It has been common to find official social network profiles of organizations and governmental entities related to the services they provide or are responsible for (water, waste, transportation, cultural events, etc.) and that are used by citizens as a gateway to directly interact and communicate their complains and problems about those services. The present work proposes to apply machine learning algorithms over the urban data generated by social networks in order to create classifiers to automatically categorize citizens messages according to the different cities services dimensions. For that, two distinct text datasets in Portuguese were collected from two social networks: Twitter (1,950 tweets) and Colab (65,066 posts). The texts were mapped according to the different ISO 37120 categories, preprocessed and mined through the use of 11 algorithms implemented in Scikit-Learn. Initial results pointed out the feasibility of the proposal with models achieving average F1-measures around 59% for F1-macro and 75% for F1-micro when using *Linear Support Vector Classification*(LSVC) and *Complement Naive Bayes*(CNB). However, as the datasets were highly unbalanced, the performances of the models vary significantly for each ISO category, with the best results occurring for *Transportation* (87%), *Energy* (83%) and *Wastewater* (74%). The classifiers generated here can be integrated on a number of different city services and systems such as: governmental support decision systems, citizens complain systems, communities dashboards, police offices, transportation's companies, cultural producers, environmental agencies, and recyclers' companies.

Keywords: Topic Classification. Machine Learning. Social Networks. Smart City Services. ISO 37120.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

SUMÁRIO

1 INTRODUÇÃO

Uma cidade inteligente pode ser definida como uma cidade de alta tecnologia com vários recursos para resolução dos problemas urbanos (SHARMA; RAJPUT, 2017). Espera-se que as cidades melhorem constantemente os serviços prestados aos seus cidadãos em termos de impacto econômico (maior eficiência) e social (efetividade para atender às necessidades e desejos de seus *stakeholders*) (AGUILERA et al., 2017). Desenvolver cidades sustentáveis e inteligentes em todo o mundo vem também como resposta a um movimento acelerado de urbanização que começou há algumas décadas. Enquanto apenas 30% da população mundial vivia nos centros urbanos em 1950, esse percentual aumentou para 55% em 2018 e está projetado para ser de 65% até 2050 (UNITED NATIONS, 2018). No Brasil, o censo de 2010 mapeou que 84% da população vivia em cidades (IBGE, 2010).

Preocupações com problemas gerados pela rápida urbanização e o uso de tecnologia para resolvê-los não são uma questão recente, como pode ser visto em Tokmakoff e Billington (1994), no entanto, tem havido um crescimento exponencial de pesquisas relacionadas à Cidades Inteligentes nos últimos anos. Isso se deve principalmente à expansão da cobertura da internet e à disseminação de tecnologias móveis, juntamente com o crescimento espacial vertiginoso das cidades e a busca pela sustentabilidade baseada na preocupação com as questões ambientais (DAMERI; COCHIA, 2013; DAMERI, 2017).

As tecnologias combinadas com os sistemas da cidade permitem um ambiente em que os mundos real e digital estão continuamente interagindo (SUZUKI, 2017), aumentando a capacidade de descobrir conhecimento, mas também trazendo vários desafios em relação à manipulação, mineração e visualização desses dados urbanos. Diversos métodos de processamento e uso de dados urbanos têm recebido atenção da comunidade científica nos últimos anos, pois há uma série de desafios relacionados, como por exemplo lidar com o tamanho e a variedade dos dados, a complexidade dos modelos físicos, as questões de segurança e privacidade individual, entre outros (JUNG, 2017).

Os dados são "o poder e a energia" de uma cidade (ANTHOPOULOS, 2017) e podem ser considerados a essência da inteligência das sociedades e economias, uma vez que a jornada de transformação começa com dados, continua através de serviços eletrônicos e, finalmente, melhora a qualidade de vida (WOJCIECH, 2013). Neste contexto, as Redes Sociais Online (RSO) são um componente chave dos dados urbanos. Os

dados das RSO aumentaram drasticamente e tornaram-se inestimáveis tanto para a academia quanto para o mundo empresarial(ZHANG et al., 2018). RSO causaram uma mudança em como as pessoas se comunicam e compartilham conhecimento (GUPTA et al., 2018; SAPOUNTZI; PSANNIS, 2018). A análise de RSO tem sido muito utilizada nas ciências sociais (pesquisas, entrevistas, questionários) anunciando assim, a ciência social computacional (SAPOUNTZI; PSANNIS, 2018). Nessa direção, muitas técnicas de aprendizado de máquina têm sido amplamente adotadas para resolver problemas nas RSO, como detecção de *spam bots*, detecção de intrusão (COSTA et al., 2019), classificação de usuários, detecção de eventos, análise de sentimentos, aprendizado de tópicos e muitos outros campos (ZHANG et al., 2018). Cada usuário de mídia social pode ser visto como um agente ou sensor que continuamente compartilha informações (GELERNTER; MUSHEGIAN, 2011) temporais (quando) e espaciais (onde), revelando atividades e opiniões sobre o ecossistema urbano. Devido à sua natureza multidimensional, é possível conectar os dados de RSO aos indicadores da cidade, facilitando o monitoramento de ocorrências de reclamações e eventos em todo o ambiente urbano.

Tem sido cada vez mais comum encontrar perfis de RSO relacionados à municípios ou à empresas responsáveis por serviços importantes na cidade, como água, lixo e transporte. A manutenção de contas de RSO faz parte da estratégia de comunicação dessas entidades, pois vários cidadãos as seguem buscando por informações que reconhecem como importantes. O mesmo acontece com algumas organizações da comunidade que criam perfis para compartilhar informações sobre clima, tráfego, eventos culturais, etc. Além de consumir as informações compartilhadas, os cidadãos geralmente interagem com essas contas quando enfrentam problemas que consideram ser de responsabilidade dessas entidades resolver. Estas interações são uma fonte rica de informações e podem ser integradas a sistemas e aplicativos para Cidades Inteligentes, a fim de monitorar melhor as diferentes dimensões da cidade e os serviços relacionados às mesmas.

Sistemas de indicadores em ambientes com muitos atores (como as cidades) facilitam a abertura do diálogo, possibilitando o compartilhamento de informações, aprendizado e consenso entre especialistas e leigos, entre governo, empresas e cidadãos, e entre diferentes níveis de governo (federal, estadual ou municipal) (HOLDEN, 2013). Ter clareza das dimensões a serem monitoradas dentro do complexo contexto da cidade é fundamental para um gerenciamento efetivo e uma comunicação mais clara entre os atores da cidade. Nessa direção, diferentes

modelos de indicadores de cidades foram desenvolvidos para ajudar a medir o desempenho das mesmas. Este é o caso da ISO 37120, um modelo desenvolvido com base no fato de que os indicadores existentes no nível local muitas vezes não são padronizados, consistentes ou comparáveis ao longo do tempo ou entre as cidades (ISO, 2014). A ISO 37120 está focada em serviços urbanos e qualidade de vida, visando contribuir para a sustentabilidade da cidade.

Este trabalho apresenta uma proposta para usar dados coletados de perfis de RSO como informações de entrada para os serviços de Cidades Inteligentes. A ideia principal é desenvolver modelos capazes de classificar automaticamente mensagens de RSO de acordo com as diferentes dimensões de serviços urbanos (dimensões ISO 37120) de maneira que possam ser integrados dentro de soluções para Cidades Inteligentes. Neste contexto, os serviços urbanos devem ser entendidos como serviços públicos, financiados ou autorizados pelo governo (água, energia, transporte público), bem como serviços privados considerados essenciais para Cidades Inteligentes (transporte privado, eventos culturais). Para testar a viabilidade da presente proposta, foram coletados dados de duas RSO distintas (Twitter e Colab) e avaliou-se o desempenho de diferentes modelos de classificação por meio de técnicas de aprendizado de máquina.

Dentro do contexto exposto, este trabalho está fundamentado sobre as seguintes perguntas de pesquisa:

PP1. Como classificar as interações de RSO nas dimensões da ISO 37120?

PP2. Quais são os desafios técnicos para estender essa classificação a outros modelos de indicadores de cidades?

PP3. Qual seria uma proposta de serviço de classificação viável e suas aplicações nas demandas da cidade?

1.1 OBJETIVOS

Considerando que a motivação deste trabalho é utilizar a informação contida nas mensagens de redes sociais criando formas de classificar as mesmas dentro de dimensões de indicadores de modelos de Cidades Inteligentes, o objetivo geral e os objetivos específicos são apresentados a seguir.

1.1.1 Objetivo Geral

Propor modelo para classificação automática de mensagens de redes sociais dentro das dimensões de Cidades Inteligentes estabelecidas pelo padrão ISO 37120.

1.1.2 Objetivos Específicos

- Analisar, definir e implementar as etapas necessárias para a construção das bases com mensagens das redes Twitter e Colab.
- Desenvolver dicionário de termos para ISO 37120.
- Determinar Modelo de Classificação Supervisionada (MCS) para ISO 37120 usando as redes Twitter e Colab:
  - Selecionar estratégias de classificação supervisionada;
  - Implementar a aplicação dos classificadores;
  - Avaliar resultados e determinar o melhor MCS;
- Propor uma Estrutura de Classificação de Mensagens da Cidade (ECMC) capaz de contemplar mais de um modelo de Cidades Inteligentes, descrevendo possíveis aplicações e potenciais consumidores do serviço.

1.2 ADERÊNCIA AO PPGTIC

O PPGTIC possui três linhas de pesquisa: Tecnologia Educacional, Tecnologia Computacional e Tecnologia, Gestão e Inovação. A pesquisa desenvolvida nesta dissertação se enquadra na linha Tecnologia Computacional.

> "O objetivo da linha Tecnologia Computacional é desenvolver modelos, técnicas e ferramentas computacionais auxiliando na resolução de problemas de natureza interdisciplinar. Especificamente, esta linha de pesquisa procura desenvolver novas tecnologias computacionais para aplicação nas áreas de educação e gestão." (PPGTIC, 2016).

Essa dissertação abrange o tema multidisciplinar das Cidades Inteligentes. A cidade pode ser vista como um conjunto de sistemas conforme representação da Figura 1, onde todos os subsistemas envolvem áreas de conhecimento diversas. As técnicas e modelos desenvolvidos neste trabalho poderão auxiliar os atores da cidade na compreensão da manifestação popular e no aprendizado das várias dimensões do ambiente urbano. O modelo proposto poderá beneficiar gestores de serviços urbanos na comunicação com os cidadãos e na compreensão da extensão dos problemas nas diversas dimensões.

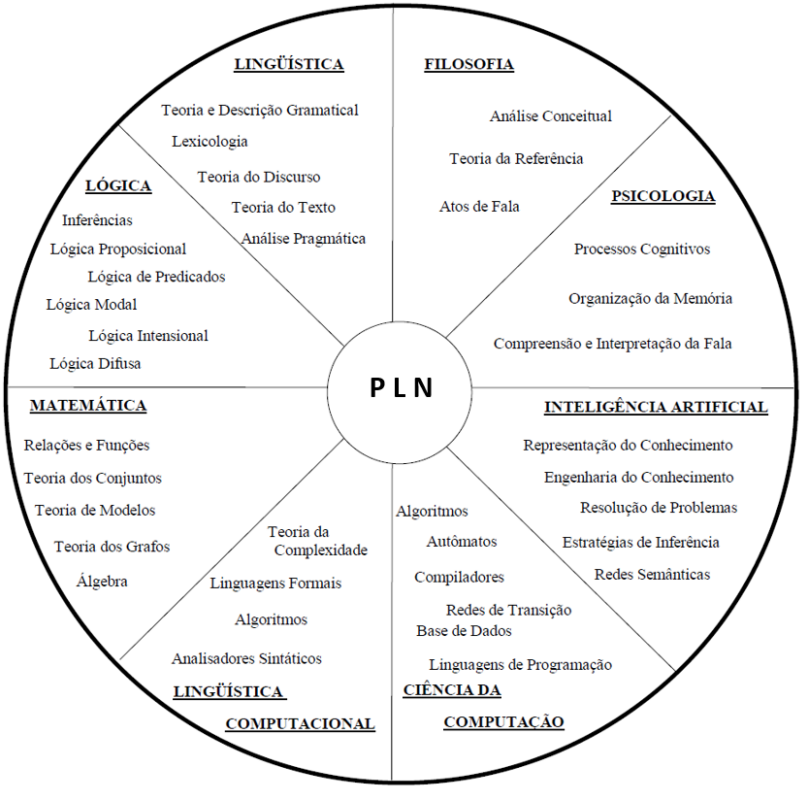
Figura 1 – A Cidade como um Sistema de Sistemas

Fonte: Extraído de Suzuki (2017).

A pesquisa desta dissertação foi desenvolvida utilizando dados de redes sociais que são atualmente uma fonte de informações muito relevante do *Big Data* Urbano. Segundo Thakuriah, Tilahun e Zell-

ner (2017), o *Big Data* Urbano está emergindo como uma importante área de pesquisa multidisciplinar. Além da utilização dos dados das redes sociais esta pesquisa investiga questões e técnicas relacionadas à PLN (Processamento de Linguagem Natural), que é sobretudo um tema multidisciplinar como é demonstrado por Silva (1996) na Figura 2.

Figura 2 – PLN - um tema multidisciplinar



Fonte: Adaptado de Silva (1996).

1.3 METODOLOGIA

O trabalho será realizado através de uma pesquisa exploratória e tecnológica com a implementação de classificadores para as mensagens

coletadas das redes Twitter e Colab dentro das dimensões do padrão ISO 37120, por meio de algoritmos disponíveis na biblioteca Scikit-learn[1]. Para atingir os objetivos, o trabalho foi dividido nas seguintes etapas:

- Realizar a revisão exploratória de literatura sobre os modelos de Cidades Inteligentes mais utilizados avaliando suas dimensões e efetuando a correlação manual dos mesmos com a norma ISO 37120.
- Explorar os trabalhos relacionados à aplicação de mídias sociais no contexto das cidades utilizando aprendizagem de máquina.
- Investigar técnicas para coleta de dados das redes sociais escopo deste trabalho.
- Explorar os fundamentos teóricos da classificação supervisionada com enfoque em processamento de textos, técnicas de seleção de atributos (*features*, em inglês) e as principais métricas de avaliação dos resultados.
- Estudar a biblioteca Scikit-learn com foco nos algoritmos de classificação supervisionada e no desenvolvimento de rotinas para a aplicação dos mesmos utilizando a linguagem Python[2].
- Investigar modelos de classificação que poderiam ser utilizados.

1.4 CONTRIBUIÇÕES DO TRABALHO

A pesquisa realizada durante o desenvolvimento desta dissertação culminou em duas publicações e uma submissão em andamento. A primeira publicação tratou de uma revisão das tecnologias utilizadas na busca por Rodovias Inteligentes (*Smart Roads*) (BENCKE; PEREZ; ARMENDARIS, 2017), abordando tendências tecnológicas a respeito da dimensão Transportes. A segunda publicação trata da revisão e comparação de modelos de indicadores para Cidades Inteligentes onde destacou-se que alguns modelos podem se aplicar melhor à realidade de determinadas cidades do que outros (BENCKE; PEREZ, 2018). A submissão em andamento aborda os experimentos de coleta e classificação dos dados do Twitter e Colab, bem como proposta para ECMC e os resultados obtidos na determinação do MCS para ISO 37120.

[1] https://scikit-learn.org/
[2] https://www.python.org/

Até o momento da finalização desta dissertação, não foram encontrados trabalhos que abordassem a classificação das mensagens de RSO nas dimensões da ISO 37120 ou que propusessem uma estrutura de serviços de classificação viável capaz de adicionar outros modelos de indicadores para cidades. Nos limites do conhecimento da autora, esta é a primeira pesquisa a vincular postagens de RSO em português a um modelo de cidade inteligente. Também não encontrou-se base de conhecimento em outros idiomas para a ISO 37120 semelhante a que foi desenvolvida em português.

1.5 ORGANIZAÇÃO DO TEXTO

Este documento está organizado em mais seis capítulos. No segundo capítulo, é realizada uma revisão da fundamentação teórica das principais áreas de pesquisa relacionadas a este trabalho, tais como, Redes Sociais, Modelos de Cidades Inteligentes, Mineração de Texto e Aprendizado de Máquina com foco na Classificação de Textos. O terceiro capítulo, apresenta trabalhos relacionados aplicando dados de RSO ao contexto da cidade: pesquisas identificando problemas urbanos, integrando esses dados à serviços urbanos ou vinculando os mesmos a modelos de indicadores para Cidades Inteligentes. No capítulo quatro, são apresentadas propostas e experimentos que visam responder às perguntas desta pesquisa e atingir os objetivos estabelecidos. No quinto capítulo, os resultados são apresentados e é feita uma análise dos mesmos em relação às perguntas da pesquisa. Finalizando, o capítulo seis apresenta as últimas considerações e as propostas de trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

Este capítulo apresenta conceitos sobre Redes Sociais e Modelos de Cidades Inteligentes, além de uma revisão de técnicas de Recuperação da Informação, Mineração e Classificação de Texto.

2.1 REDES SOCIAIS

Com a tecnologia e o estilo de vida em constante mudança, as RSO desempenham um papel importante na vida de todos os usuários (GUPTA et al., 2018). RSO, comunidades web, blogs, Wikipedia e outras formas de mídia colaborativa online facilitaram a criação de conteúdo original, ideias e opiniões, conectando milhões de pessoas em toda a World Wide Web, de maneira financeira e trabalhista (HUSSAIN; CAMBRIA, 2018). A mídia social é uma fonte de dados muito relevante para o *Big Data Social*, que se refere à análise de grandes quantidades de dados que podem vir de várias fontes distribuídas, mas com um forte foco nas mídias sociais (BELLO-ORGAZ; J.JUNG; CAMACHO, 2015).

As autoridades podem usar os dados urbanos para construir uma variedade de informações automaticamente e desenvolver serviços inteligentes que melhorem a vida cotidiana dos cidadãos, economizem recursos ambientais ou ajudem no enfrentamento de emergências (PANAGIOTOU et al., 2016). O Twitter tem sido objeto de muitas pesquisas e tem desafiado a academia, especialmente devido à sua natureza informal. Os *tweets* são mensagens informais curtas, geralmente abrangendo uma frase ou menos, tendem a ter muitos erros de ortografia, termos de gíria, formas abreviadas de palavras e também podem ter marcadores especiais (KIRITCHENKO; ZHU; MOHAMMAD, 2014). Embora muitos esforços científicos tenham sido empregados trazendo progresso em direção as subtarefas específicas de análise de redes sociais, derivar conhecimento de dados das RSO continua sendo um grande desafio, uma razão importante para isso é a troca de dados em linguagem coloquial com grande quantidade de ruídos (SAPOUNTZI; PSANNIS, 2018).

2.1.1 Twitter

O Twitter[1] é uma RSO de grande popularidade, sendo reconhecida como uma fonte importante de notícias e que produz vasto volume de informações públicas (WEILER; GROSSNIKLAUS; SCHOLL, 2011; BELLO-ORGAZ; J.JUNG; CAMACHO, 2015; CARVALHO et al., 2017). Hoje conta com cerca de 330 milhões de usuários ativos e, segundo Carvalho et al. (2017), tornou-se uma ferramenta importante também para disseminar posições e ideias, e para comentar e analisar eventos atuais do mundo.

É um RSO extremamente dinâmica e muito conectada com o espaço urbano: é comum ver *tweets* de usuários sobre o que está acontecendo ao seu redor, como opiniões a respeito de eventos, reclamações sobre problemas no meio ambiente, etc. Vários governos locais têm mantido perfis no Twitter e em outras redes sociais como forma de interagir com os cidadãos e publicar informações, campanhas, alertas, etc.

2.1.2 Colab

O Colab[2] é uma RSO brasileira focada em promover discussões sobre o espaço urbano, com o objetivo claro de coletar expressões de cidadãos sobre questões da cidade. O usuário se inscreve gratuitamente, conecta-se a uma cidade e pode postar problemas ou sugestões relacionadas a esse local. O *feed*[3] dos usuários mostra os problemas relatados sobre a cidade com a qual estão conectados, no entanto, eles podem mudar de cidade quando necessário. As mensagens podem ser reforçadas por outros usuários, bem como compartilhadas em outras redes sociais (como o Twitter).

A empresa se mantém através da parceria com alguns governos municipais, atuando como interlocutora na identificação de problemas e também para comunicar as soluções. No entanto, a postagem de mensagens pelo usuário é totalmente gratuita e independente do fato de a empresa ter feito parceria com uma cidade específica. Os usuários são convidados a anexar uma foto da questão urbana a que se referem e a descrever o problema, classificando-o dentro das várias categorias internas de problemas urbanos existentes. Na Figura 3 encontram-se

[1] https://twitter.com/
[2] https://www.colab.re/
[3] Lista atualizada constantemente com postagens de perfis que o usuário segue.

Figura 3 – Exemplos de postagens Colab

Fonte: Extraído da Rede Social Colab.

imagens de postagens.

2.2 SMART CITY

Não há uma definição única para *Smart City* (Cidade Inteligente), sendo possível verificar algumas descrições utilizadas nas pesquisas abordadas por Cocchia (2014). O que a maioria das definições de Cidades Inteligentes têm em comum é que elas consideram o uso de tecnologias e dos dados como meios para resolver os desafios econômicos, sociais e ambientais da cidade. Uma definição que considera as questões de interoperabilidade e integração dos dados, mas também mantém o cidadão no centro é a seguinte:

> "Uma Cidade Inteligente é uma cidade em que seus aspectos sociais, empresariais e tecnológicos são suportados pelas Tecnologias de Informação e Comunicação para melhorar a experiência do cidadão dentro da cidade. Para conseguir isso, a cidade oferece serviços públicos e privados que operam de forma integrada, acessível e sustentá-

vel". (SANTANA et al., 2017)

Vários modelos conceituais são encontrados na literatura para definir Cidades Inteligentes. Uma comparação pode ser verificada em Anthopoulos (2015) onde os modelos conceituais são sintetizados em oito componentes principais das Cidades Inteligentes:

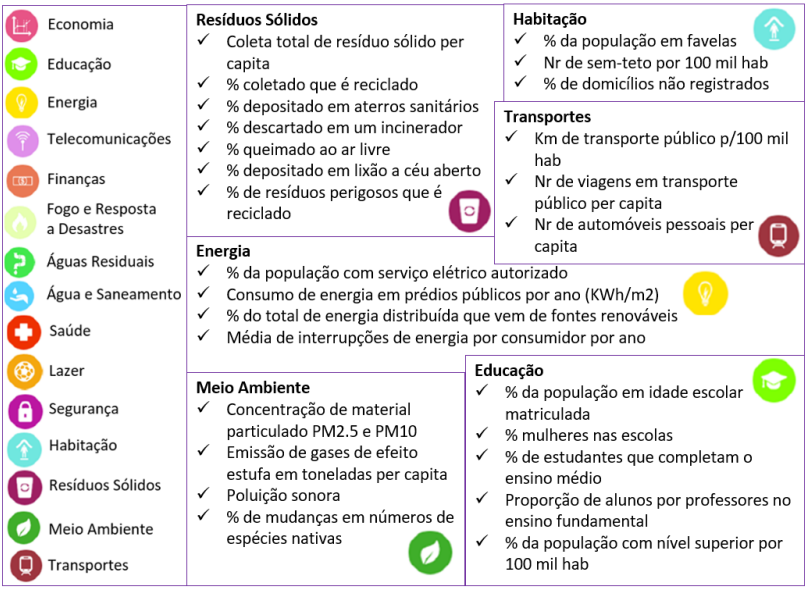
- *Smart Infrastructure*: redes de água, esgoto, energia, ruas, prédios, etc. dotados de tecnologia (sensores, *smart grids*, etc).
- *Smart Transportation* ou *Smart Mobility*: sistemas de transporte dotados de monitoramento e controle em tempo real.
- *Smart Environment*: inovação e incorporação das Tecnologias da Informação e Comunicação (TIC) para proteção e gestão dos recursos naturais (sistemas de gestão do lixo, controle das emissões, reciclagem, sensores para monitoramento da poluição, etc.).
- *Smart Services*: utilização das TIC nas diversas áreas de serviços como saúde, educação, turismo, segurança, controle da demanda e tempos de resposta.
- *Smart Governance*: uso da tecnologia para a gestão dos serviços, participação e engajamento dos cidadãos.
- *Smart People*: cidadãos conscientes, criativos e colaborativos.
- *Smart Living*: inovações relacionadas a melhoria na qualidade de vida no espaço urbano.
- *Smart Economy*: TIC fortalecendo o desenvolvimento dos negócios, o emprego e o crescimento urbano.

2.3 MODELOS PARA CIDADES INTELIGENTES

A ISO desenvolveu o padrão 37120 com o objetivo de fornecer um conjunto de indicadores como uma recomendação do que medir e como deve ser medido na gestão das Cidades Inteligentes (ISO, 2014). O principal objetivo do desenvolvimento deste modelo é ajudar as cidades a medir a gestão do desempenho dos serviços municipais e a qualidade de vida ao longo do tempo, facilitar a aprendizagem de uma cidade com a outra, permitir a comparação em uma ampla gama de medidas e compartilhar melhores práticas. O portal do *World Council on*

City Data (WCCD[4], Conselho Mundial de Dados Municipais) disponibiliza dados de todas as cidades que aderem à ISO 37120. Este portal fornece às cidades uma base confiável de dados globalmente padronizados que ajudarão no desenvolvimento de conhecimento básico para comparações em nível global. A ISO 37120 estabelece um conjunto de temas/dimensões relacionados aos serviços da cidade e à qualidade de vida, onde cada tema possui um conjunto de indicadores. Exemplos de alguns indicadores ISO podem ser visualizados na Figura 4.

Figura 4 – Exemplos de Indicadores da ISO 37120



Fonte: Adaptado de ISO (2014).

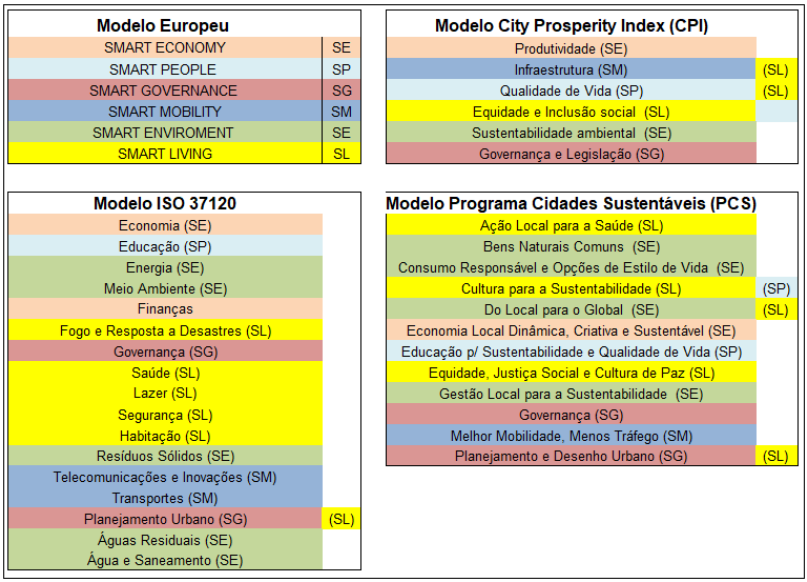
Existem outros modelos, como o *City Prosperity Initiative* (CPI[5], Iniciativa para Prosperidade da Cidade), que se autodeclara baseado nos direitos humanos. Modelos de indicadores de cidades geralmente têm um bom nível de correspondência e dividem a cidade em dimensões semelhantes. Para determinar o melhor modelo para um local, é importante avaliar se ele possui dimensões que traduzam adequadamente a realidade da cidade (BENCKE; PEREZ, 2018). Por exemplo, cidades com índices muito bons de desenvolvimento humano podem ter maior

[4] http://open.dataforcities.org/
[5] http://cpi.unhabitat.org

interesse em questões ambientais ou em soluções tecnológicas de ponta, diferentes de lugares com altos índices de desigualdade, com várias discussões sobre consciência social ou necessidades humanas básicas.

Bencke e Perez (2018) analisam quatro modelos de indicadores usados para classificar cidades e permitir monitoramento da evolução das mesmas em direção às cidades sustentáveis, inteligentes e humanas. Os modelos ISO 37120, CPI, o modelo europeu da Universidade de Viena (TUWIEN, 2015) e o Programa Cidade Sustentável do Brasil[6] são comparados segundo seus objetivos, escopo, metodologia e estratégias de disseminação. Além disso, é efetuada comparação das dimensões dos modelos, apresentada na Figura 5. Demonstra-se por cores as categorias equivalentes usando como diretriz o modelo europeu.

Figura 5 – Dimensões dos Modelos de Cidades Inteligentes



Fonte: Extraído de Bencke e Perez (2018).

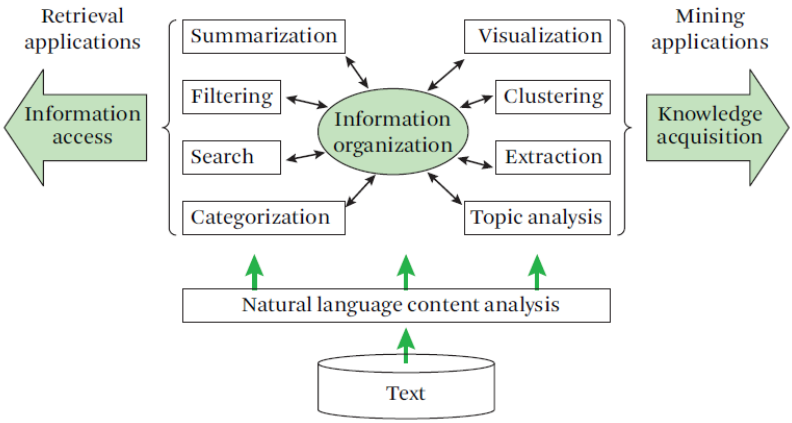
[6] www.cidadessustentaveis.org.br

2.4 MINERAÇÃO DE TEXTO

A mineração de textos ou descoberta de conhecimento no texto (*Knowledge Discovery from Text*, KDT) se refere ao processo de extrair informação de alta qualidade de textos estruturados, semiestruturados e não estruturados (ALLAHYARI et al., 2017). A aquisição de conhecimento a partir de texto é frequentemente obtida através do processo de mineração de texto, que pode ser definido como extrair dados do texto para descobrir conhecimento útil (ZHAI; MASSUNG, 2016). Existem diversas atividades inseridas na mineração de texto incluindo extração da informação, Processamento de Linguagem Natural (PLN) e aprendizagem de máquina.

Aggarwal e Zhai (2012a) destacam que o objetivo do acesso à informação é conectar as informações corretas com os usuários certos no momento adequado, com menor ênfase no processamento ou na transformação das informações de texto. Para os autores a mineração de texto vai além do acesso à informação, pretende também auxiliar os usuários a analisar informações, facilitando a tomada de decisões. Na Figura 6, Zhai e Massung (2016) apresentam um *framework* conceitual para Sistema de Informação de Texto (SIT) contendo vários módulos suportados pela análise de conteúdo baseada em técnicas de PLN. Este módulo de análise de conteúdo permite que o SIT transforme texto bruto não estruturado em representações com mais significado.

Figura 6 – Framework Conceitual para SIT



Fonte: Extraído de Zhai e Massung (2016).

2.4.1 Processamento de Linguagem Natural

O Processamento de Linguagem Natural (PLN) tem como principal meta construir modelos computacionais que possibilitem a comunicação entre homem e máquina, de forma efetiva, via linguagem natural (CONTERATTO, 2015). Para Zhai e Massung (2016) a PLN é um componente fundamental dos SIT, pois a eficácia do sistema em ajudar os usuários a acessar e analisar dados de texto é amplamente determinada pelo quão bem o sistema pode entender o conteúdo dos dados de texto. Os autores listam como principais tarefas da PLN as análises: léxica, sintática, semântica, pragmática e de discurso. Para um bom resultado nestas análises são necessárias algumas tarefas básicas de processamento que segundo Uysal e Gunal (2014) são importantes para o bom desempenho dos SIT: tokenização, filtragem, *stemming* e lematização.

Dada uma sequência de caracteres e uma unidade de documento definida, a tokenização é a tarefa de dividi-la em partes (tokens) (MANNING; RAGHAVAN; SCHüTZE, 2009). Muitos conceitos técnicos, nomes de organizações e produtos são compostos por múltiplas palavras. Para processar consultas sobre conceitos compostos de duas ou mais palavras Manning, Raghavan e Schütze (2009) sugerem um índice de frases. Para isso é necessário um índice para cada número de palavras a serem consideradas juntas (n-gramas). Tokenizar os documentos considerando n-gramas gera um número maior de tokens mas possibilita a identificação de determinadas expressões que podem facilitar o processo de classificação do texto ao trazerem mais significado para as palavras juntas do que separadas. Um exemplo de tokenização 1, 2 e 3-gramas para a frase *"tem invasão na área de preservação"*:

- 1-grama:[tem] [invasão] [na] [área] [de] [preservação]
- 2-gramas: [tem invasão] [invasão na] [na área] [área de] [de preservação]
- 3-gramas: [tem invasão na] [invasão na área] [na área de] [área de conservação]

A tarefa de filtragem geralmente é aplicada nos documentos do corpus para remover palavras que não tenham muito conteúdo denominadas *stopwords*, por exemplo, preposições, conjunções, etc. (ALLAHYARI et al., 2017). Segundo Manning, Raghavan e Schütze (2009), o objetivo do *stemming* e da lematização é o mesmo: reduzir as formas flexionadas e, por vezes, as formas relacionadas derivadas de uma

palavra, para uma forma básica comum. Os autores mencionam que o *stemming* normalmente é um processo heurístico mais grosseiro que corta as extremidades das palavras na esperança de atingir esse objetivo, enquanto a lematização faz uso de um vocabulário e efetua análise morfológica das palavras, buscando remover somente terminações flexionais e retornar a forma básica da palavra ou a "forma usada no dicionário", que é conhecida como o lema.

O interesse em investigar algoritmos de *stemming* começou na década de 1960 e ainda está em andamento. Enquanto os primeiros esforços foram dedicados a criar *stemmers* baseados em regras, hoje em dia o foco da pesquisa é em *stemmers* estatísticos, que não exigem nenhum conhecimento linguístico acerca da linguagem para a qual eles são projetados (FLORES; MOREIRA, 2016). De acordo com Santos (2015), a correta grafia das palavras é importante para o bom funcionamento dos *stemmers*, incluindo a acentuação. Para isso a aplicação de corretores é recomendada.

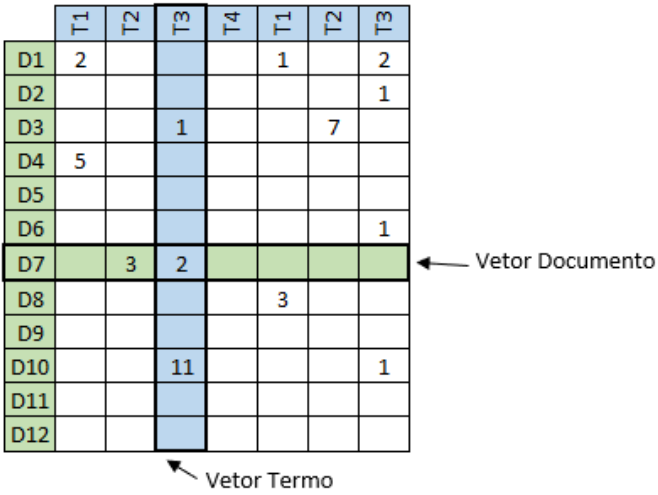
2.4.2 Modelos de Representação

Baeza-Yates e Ribeiro-Neto (2013) classificam os modelos clássicos de Recuperação da Informação (RI) em três principais: o modelo Booleano (onde os termos de indexação não têm peso associado) e os modelos vetorial e probabilístico (os termos de indexação possuem pesos associados com o objetivo de melhorar o ordenamento dos documentos).

No modelo vetorial tanto os documentos como a consulta são vistos como vetores t-dimensionais. Desta forma, localizar documentos em um corpus D que satisfaçam uma consulta $\vec{q}$ se trata de comparar o vetor documentos $\vec{d}$ a $\vec{q}$ (SALTON; WONG; YANG, 1975; SALTON; YANG; YU, 1975; BAEZA-YATES; RIBEIRO-NETO, 2013),

A ocorrência de um termo em um documento estabelece uma relação entre eles que pode ser quantificada de diversas maneiras: simplesmente computando a presença do termo ou não, comumente chamado de *Bag Of Words* (BOW) ou matriz termo-documento, ou adicionando mais informação a BOW como a frequência de ocorrência do termo ou outras formas de ponderação (por exemplo, TF-IDF) (BAEZA-YATES; RIBEIRO-NETO, 2013) como pode ser observado na Figura 7.

Figura 7 – Matriz Termo-Documento



Fonte: Elaborado pela autora.

2.4.2.1 TF-IDF

O esquema de ponderação denominado *Term Frequency–Inverse Document Frequency* (TF-IDF) atribui um peso *tf-idf* a um termo t em um documento d dado pela equação 2.1 (MANNING; RAGHAVAN; SCHüTZE, 2009). Existem diversas variações do TF-IDF e algumas delas são apresentadas por Baeza-Yates e Ribeiro-Neto (2013) na Figura 8.

$$tf\text{-}idf_{t,d} = tf_{t,d} \times idf_t \tag{2.1}$$

Segundo Manning, Raghavan e Schütze (2009) o *tf-idf*:

- aumenta muito quando t ocorre várias vezes dentro de um pequeno número de documentos (concedendo assim alto poder discriminatório a esses documentos);
- diminui quando t ocorre menos vezes em um documento, ou ocorre em muitos documentos (oferecendo assim um sinal de relevância menos pronunciado);
- diminui muito quando t ocorre em praticamente todos os documentos.

Figura 8 – Variantes de TF e IDF

Esquema de ponderação	Peso TF
binário	{0,1}
frequência bruta	$f_{i,j}$
normalização logarítmica	$1 + \log f_{i,j}$
normalização dupla 0,5	$0{,}5 + 0{,}5\frac{f_{i,j}}{max_i f_{i,j}}$
normalização dupla K	$K + (1-K)\frac{f_{i,j}}{max_i f_{i,j}}$
Esquema de ponderação	**Peso IDF**
unário	1
frequência inversa	$\log \frac{N}{n_i}$
frequência inversa suave	$\log(1 + \frac{N}{n_i})$
frequência inversa máxima	$\log(1 + \frac{max_i n_i}{n_i})$
frequência inversa probabilística	$\log \frac{N-n_i}{n_i}$

Fonte: Extraído de Baeza-Yates e Ribeiro-Neto (2013).

2.4.3 Classificação de Textos

Classificar textos em categorias enriquece a representação do texto e possibilita análises mais efetivas e profundas (ZHAI; MASSUNG, 2016). Os tipos de classificação podem ser *multilabel* (também conhecida como multirrótulo, *multivalue* ou *any-of*), onde um documento pertence a mais de uma classe, ou *multiclass* (rótulo-único, *multinomial* ou *one-of*), cada documento pertence a uma só classe (MANNING; RAGHAVAN; SCHüTZE, 2009; BAEZA-YATES; RIBEIRO-NETO, 2013).

O problema de classificação com categorias predefinidas, busca determinar a que classe(s) um determinado objeto pertence. Manning, Raghavan e Schütze (2009) formalizam este problema no contexto da classificação de textos da seguinte forma: dada uma coleção de documentos $\mathbb{X} = \{d_1, d_2 ... d_k\}$ e um conjunto de classes $\mathbb{C} = \{c_1, c_2 ... c_L\}$, deseja-se aprender um classificador γ que identifique a que classe os documentos pertencem, conforme descrito pela Equação 2.2.

$$\gamma : \mathbb{X} \rightarrow \mathbb{C} \tag{2.2}$$

Dependendo do mecanismo de aprendizado usado os algoritmos podem ser basicamente de dois tipos: supervisionado (há uma base rotulada de treinamento) ou não supervisionado (não há exemplos de treinamento). Na classificação supervisionada para determinar o classificador da Equação 2.2 é fornecido o conjunto de treinamento D com dados rotulados $\langle d, c\rangle$, onde $\langle d, c\rangle \in \mathbb{X} \times \mathbb{C}$ (MANNING; RAGHAVAN; SCHüTZE, 2009).

2.4.3.1 Classificação de Texto nas RSO

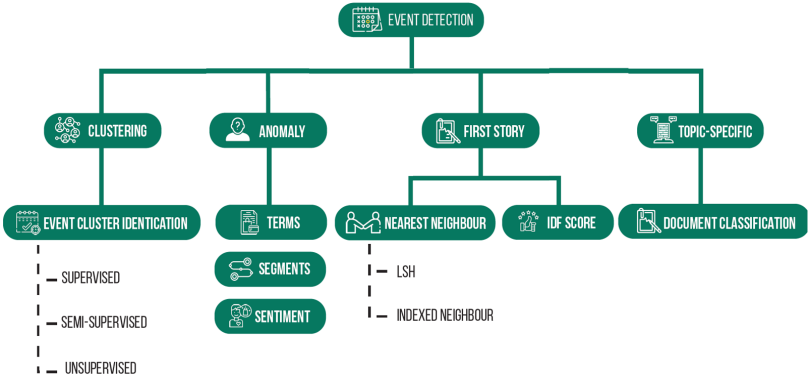
Muitos trabalhos que lidam com problemas de classificação para dados oriundos de RSO geralmente utilizam abordagens denominadas Detecção de Tópicos (*Topic Detection*, TD) , Detecção de Eventos (*Event Detection*, ED) ou Classificação de Tópicos (*Topic Classification*, TC). De acordo com Milioris (2018), Detecção de Tópicos (ou sensores de tendência) é o problema de identificar automaticamente tópicos de grandes coleções de documentos e extrair o tema principal sem qualquer intervenção humana; enquanto a Classificação de Tópicos (ou Classificação de Texto) é a tarefa de atribuir rótulos ou categorias predefinidas a textos ou documentos.

Fiscus e Doddington (2002) definem um Evento como "algo que acontece em algum momento e local específicos, juntamente com todas as pré-condições necessárias e consequências inevitáveis". A detecção de eventos envolve alta complexidade e impacto social e abrange muitos desafios, incluindo o processamento de grandes volumes de dados e altos níveis de ruído. De acordo com Panagiotou, Katakis e Gunopulos (2016), a maioria das abordagens técnicas para detecção de eventos, pelo menos em um primeiro estágio, utilizam *clustering* (agrupamento). Todavia, ED deve ser entendido como um problema multidimensional e pode ser abordado de uma das maneiras descritas na taxonomia da Figura 9.

Para responder à PP1, será necessária uma abordagem de Classificação de Tópicos, pois haverá classes fixas predefinidas (as dimensões do modelo de Cidade Inteligente). A solução proposta tem que lidar com expressões de cidadãos que podem ser postadas por apenas um indivíduo. Com base nesses fatos, a Detecção de Eventos ou Detecção de Tópicos não será adequada num primeiro momento, embora possam adicionar valor para determinar prioridades de postagens já classificadas em dimensões de cidades em etapas futuras.

Um exemplo de classificação de tópicos em RSO, não direta-

Figura 9 – Taxonomia para Detecção de Eventos



Fonte: Adaptado de Panagiotou, Katakis e Gunopulos (2016).

mente relacionado à Cidades Inteligentes, é encontrado em Wang et al. (2017). Os autores efetuam a categorização de *tweets* em 14 classes (álcool e drogas, informações sobre a família, entre outros). O problema é abordado com o método TF-IDF e melhoram os resultados usando termos que constam nas preferências de tópicos dos usuários. Os resultados do experimento mostraram que, com as preferências de tópicos, a precisão da classificação aumentará, incluindo o desempenho de cada categoria.

2.4.3.2 Seleção de Atributos

O termo "seleção de atributos" (*feature selection*) é usado no aprendizado de máquina para o processo de seleção de um subconjunto de atributos (características) usadas para representar os dados. A seleção de atributos na mineração de texto é usada principalmente em conexão com a aplicação de métodos de aprendizado de máquina e estatística do texto, sendo usado em tarefas como agrupamento ou classificação de documentos (MLADENIC, 2010). Embora a seleção de atributos também seja desejável em outras tarefas de classificação, ela é especialmente importante na classificação do texto devido à alta dimensionalidade dos termos e existência de muitos atributos irrelevantes (ruído) (AGGARWAL; ZHAI, 2012b). Neste trabalho, definiremos como "atributo" ou "característica" um termo (palavra ou expressão) que tem significado relevante para o documento em questão e para a classe que

o mesmo representa.

De maneira geral, existem dois tipos de redução de atributos: (1) Extração e/ou Transformação de atributos, concentrando-se na redução de dimensionalidade por transformação e/ou projeção de todos os atributos em subconjuntos e (2) Seleção de atributos para identificar as características importantes e remover as irrelevantes e, dessa forma, melhorar o desempenho e a precisão do algoritmo de aprendizado de máquina (AGARWAL; MITTAL, 2014).

As técnicas de seleção de atributos podem ser agrupadas de maneira geral em abordagens que são dependentes do classificador (métodos *wrapper* e *embedded*) e independentes do classificador (métodos *filter*) (BROWN et al., 2012; AGARWAL; MITTAL, 2014; LIU et al., 2017) .

- Método *filter*: todos os atributos são tratados independentemente uns dos outros e classificados de acordo com sua importância, a qual é calculada com uso de alguma função (AGARWAL; MITTAL, 2014). Os métodos *filter* executam uma análise estatística sobre o espaço de atributos para selecionar um subconjunto discriminativo de recursos (LABANI et al., 2018).

- Método *embedded*: o processo acontece explorando a estrutura de classes específicas de modelos de aprendizado para guiar o processo de seleção de atributos (BROWN et al., 2012).

- Método *wrapper*: o critério de seleção é o desempenho do preditor (algoritmo de classificação empregado), ou seja, o preditor é agrupado em um algoritmo de busca que encontrará um subconjunto de atributos que fornece o mais alto desempenho do preditor (CHANDRASHEKAR; SAHIN, 2013). As abordagens por *wrapper* oferecem vantagens significativas na generalização, embora tenham a desvantagem de um considerável custo computacional, e podem produzir subconjuntos que são excessivamente específicos (*overfitting*) para o classificador usado (BROWN et al., 2012).

2.4.3.3 Avaliação da Classificação

A avaliação de classificadores envolve a divisão aleatória de um conjunto previamente rotulado em duas partes, um conjunto de treinamento e um conjunto de teste (JAMES et al., 2013). O modelo de classificação é ajustado por meio do conjunto de treinamento e, depois

disso, é usado para predizer as classificações das observações no conjunto de teste, a partir deste resultado é possível calcular métricas de avaliação do modelo utilizado. A divisão da base apresenta o dilema de qual o percentual correto a utilizar. O ideal seria ter o maior número de amostras para treinamento, mas também ter uma base de teste estatisticamente significativa para trazer uma avaliação acurada (RUSSELL; NORVIG, 2013).

O problema descrito se amplifica no caso de bases rotuladas menores. Uma forma de endereçá-lo é utilizar o método de avaliação denominado *Cross Validation* (CV, validação cruzada). Utilizam-se todos os casos da base de dados tanto para teste quanto para treinamento, porém não ao mesmo tempo (CECHINEL; CAMARGO, 2018). Os autores esclarecem a estratégia, que consiste em dividir o conjunto de dados em n partições (*n-fold*) de igual tamanho (ou tamanhos similares), sendo que a partição n é utilizada para teste e as demais partições são utilizadas para treinamento. A ideia do CV é que cada $fold$ sirva como treino e como teste e, ao final, é efetuada a média dos resultados obtidos nos testes (RUSSELL; NORVIG, 2013), o que é melhor do que aplicar treino e testes em duas partes maiores e distintas.

Segundo Cechinel e Camargo (2018) quando a quantidade de dados é extremamente pequena, utiliza-se a CV denominada *leave-one-out* e que consiste em utilizar partições de apenas um único elemento para teste, ou seja, o número de partições gerado é igual ao número de casos da base de dados. A exatidão do modelo é calculada medindo a exatidão na predição da amostra de teste, e a exatidão final do modelo é dada pela média de todos os n experimentos (CECHINEL; CAMARGO, 2018). Este procedimento está disponível na Scikit-learn através do validador *LeaveOneOut*. Além deste, destacam-se na biblioteca os validadores *KFold* e *StratifiedKFold*. *KFold* divide a base em k *folds* e aplica o classificador em $k - 1$ folds, usando a partição deixada fora como teste. O validador *StratifiedKFold* é uma variação do *Kfold* retornando partições estratificadas, ou seja, cada conjunto contém aproximadamente o mesmo percentual de amostras por classe alvo que a base completa, o que é indicado em caso de classes muito desbalanceadas (SCIKIT-LEARN, 2019).

As métricas comumente utilizadas em sistemas de RI são Acurácia (A), Precisão (P), Revocação (R) e F1-score ($F1$), e estão apresentadas nas equações 2.3, 2.4, 2.5 e 2.6, respectivamente. Estas métricas são baseadas nos conceitos listados a seguir e a maneira como são calculadas em classificações *multiclass* pode ser visualizada na matriz da Figura 10 (SEBASTIANI, 2002):

- *True Positive* (TP, Verdadeiro Positivo): a classificação manual enquadrou o documento na classe c_i e assim também predisse o classificador.
- *True Negative* (TN, Verdadeiro Negativo): a classificação manual não enquadrou na classe c_i, nem o classificador.
- *False Positive* (FP, Falso Positivo): a classificação manual não enquadrou na classe c_i, entretanto o classificador predisse o documento como c_i.
- *False Negative* (FN, Falso Negativo): a classificação manual enquadrou o documento como c_i, entretanto o classificador não predisse o documento como c_i.

Figura 10 – Matriz de contingência para categoria c_i

Categoria c_i		**Rótulo Manual**	
		+	**-**
Classificador Automático	**+**	TP_i	FP_i
	-	FN_i	TN_i

Fonte: Adaptado de Sebastiani (2002).

$$A = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.3}$$

$$P = \frac{TP}{TP + FP} \tag{2.4}$$

$$R = \frac{TP}{TP + FN} \tag{2.5}$$

$$F1 = 2 * \frac{P * R}{P + R} \tag{2.6}$$

Em classificações *multiclass* o desempenho do modelo será avaliado pelas médias das métricas mencionadas. A forma de cálculo destas médias pode variar. De acordo com Liu, Loh e Sun (2009), a macro

média (M) considera pesos iguais às pontuações geradas de cada categoria individual, enquanto a micro média (μ) tende a ser dominada pelas categorias com amostras de treinamento mais positivas. Isso ocorre porque micro médias são obtidas pela soma de todos os resultados individuais, como pode ser observado em Equações 2.7, 2.8 e 2.9.

$$P^{\mu} = \frac{\sum_{i=1}^{|c|} TP_i}{\sum_{i=1}^{|c|} (TP_i + FP_i)} \tag{2.7}$$

$$R^{\mu} = \frac{\sum_{i=1}^{|c|} TP_i}{\sum_{i=1}^{|c|} (TP_i + FN_i)} \tag{2.8}$$

$$F1^{\mu} = 2 * \frac{(P^{\mu} * R^{\mu})}{(P^{\mu} + R^{\mu})} \tag{2.9}$$

O método das macro médias (M) primeiro avalia "localmente" as pontuações para cada categoria (c) e, em seguida, "globalmente" calcula a média dos resultados das diferentes categorias (SEBASTIANI, 2002):

$$P^{M} = \frac{\sum_{i=1}^{|c|} P_i}{|C|} \tag{2.10}$$

$$R^{M} = \frac{\sum_{i=1}^{|c|} R_i}{|C|} \tag{2.11}$$

$$F1^{M} = 2 * \frac{(P^{M} * R^{M})}{(P^{M} + R^{M})} \tag{2.12}$$

Segundo Sebastiani (2002), em um contexto de classificação de textos, considerando um corpus D, a generalidade de uma categoria c_i se refere ao percentual de documentos em D que pertencem a c_i. Os dois métodos de avaliação, $F1^{M}$ e $F1^{\mu}$, podem apresentar resultados bastante distintos, principalmente se as categorias tiverem uma generalidade muito diferente. Nesse caso, a capacidade de um classificador se comportar bem também em categorias com baixa generalidade (poucas amostras dentro da base rotulada) será enfatizada mais pela macro média (*macro average*) e muito menos pela micro média (*micro average*) (SEBASTIANI, 2002).

2.5 ALGORITMOS DE CLASSIFICAÇÃO SUPERVISIONADA

De acordo com Mitchell (1997) o aprendizado de máquina (*machine learning*) é o estudo de algoritmos que melhoram por meio da experiência:

> "Diz-se que um programa de computador aprende com a experiência E relacionada a alguma classe de tarefas T com medida de desempenho P, se seu desempenho em tarefas T, como medido por P, melhora com a experiência E.", (MITCHELL, 1997).

Existem vários algoritmos de aprendizagem de máquina e diversas formas de agrupá-los, inclusive considerando casos que se enquadram em mais de uma categoria. Vluymans (2019) discute algumas abordagens de classificação supervisionada amplamente utilizadas e as categoriza de acordo com a lógica de predição usada. Os grupos descritos pela autora estão resumidos nos tópicos das subseções seguintes. Adicionou-se as abordagens de SCIKIT-LEARN (2019), mencionando alguns algoritmos que a biblioteca disponibiliza em cada grupo, bem como a visão de outros autores que quando citados estão referenciados.

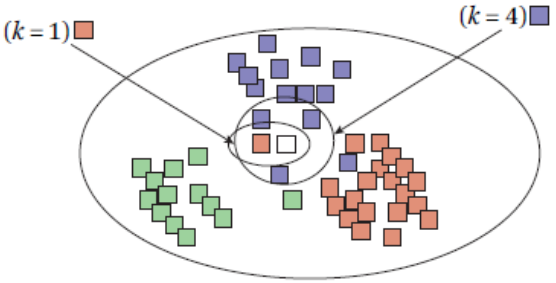
2.5.1 Vizinhos mais próximos

Este algoritmo, em inglês *k-Nearest Neighbors* (k-NN), é considerado um "aprendiz" preguiçoso, pois não constrói um modelo de classificação explícito. Em vez disso, todos os elementos de treinamento são armazenados na memória (protótipos). Para classificar um novo elemento x, são determinados os k elementos mais próximos entre os protótipos armazenados. A classe que aparece com mais frequência entre esses elementos vizinhos é predita para x.

A ideia básica por trás do k-NN é encontrar os documentos mais semelhantes ao de consulta e usar o rótulo de classe mais comum entre eles. A suposição é que documentos semelhantes terão o mesmo rótulo de classe (ZHAI; MASSUNG, 2016). Na Figura 11 verifica-se que dependendo do k usado a predição da classe do elemento vazado se modifica.

Na Scikit-learn, o algoritmo *KNeighborsClassifier* implementa o aprendizado com base nos vizinhos mais próximos de cada ponto de consulta, onde k é um valor inteiro especificado pelo usuário.

Figura 11 – Vizinhos mais próximos

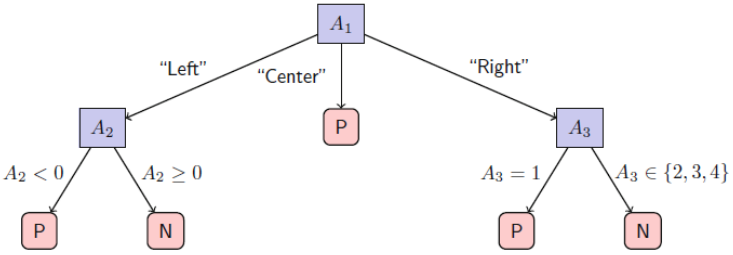


Fonte: Extraído de Zhai e Massung (2016).

2.5.2 Árvores de Decisão ou Classificação

Uma árvore é um grafo acíclico conectado. Cada nó interno corresponde a um teste baseado em um dos recursos de entrada. Os nós de folha, que formam pontos finais em caminhos a partir da raiz, representam as atribuições de classe, como o exemplo da Figura 12. Para classificar um novo elemento, ele é passado pela árvore, iniciando no nó raiz e seguindo o caminho correto em direção a uma folha. A classe vinculada à folha é atribuída ao elemento de teste.

Figura 12 – Árvore de Decisão



Fonte: Extraído de Vluymans (2019).

Na Scikit-learn o algoritmo *DecisionTreeClassifier* utiliza o classificador CART, acrônimo de *Classification And Regression Trees*, (BREIMAN et al., 1984) propondo uma versão otimizada que exige a conversão de variáveis de categoria em números. CART constrói uma árvore bi-

nária usando os atributos e calculando o limiar com maior ganho de informação para cada nodo (SCIKIT-LEARN, 2019).

2.5.3 Probabilísticos

Os preditores probabilísticos modelam o relacionamento probabilístico entre atributos preditivos e atributos alvo. São categorizados em discriminativos, modelam a distribuição de probabilidade condicional P(Y/X) (dado X, retornam a distribuição de probabilidade para Y), e generativos, modelam a distribuição de probabilidade conjunta P(X,Y).

Os algoritmos probabilísticos do tipo Naïve Bayes aplicam o teorema de Bayes que calcula a probabilidade condicional $P(c|E)$ que representa a probabilidade de uma amostra $E = (x_1, x_2, ..., x_n)$ pertencer a classe c, descrito pela equação 2.13 (ZHANG, 2004). Estes algoritmos efetuam a suposição "ingênua" de independência condicional entre cada par de atributos, dado o valor da variável de classe (SCIKIT-LEARN, 2019). Na classificação de textos estes preditores calculam para cada par documento-classe $[d_j, c_p]$ uma probabilidade, atribuindo à d_j às classes com as maiores estimativas de probabilidade (BAEZA-YATES; RIBEIRO-NETO, 2013).

$$P(c|E) = \frac{P(E|c)P(c)}{P(E)} \tag{2.13}$$

Na Scikit-learn o algoritmo *MultinomialNB* implementa Naïve Bayes onde os dados são tipicamente representados como contagens vetoriais de palavras e as probabilidades calculadas para cada documento obedecem a distribuição multinomial (SCIKIT-LEARN, 2019). A biblioteca também disponibiliza o algoritmo *ComplementNB* proposto por Rennie et al. (2003) que trata fragilidades do *MultinomialNB*, por exemplo quando uma classe tem mais exemplos de treinamento do que outra são selecionados pesos baixos para o limite de decisão. Para equilibrar a quantidade de exemplos de treinamento usados por estimativa, foi introduzida a formulação da "classe de complemento". Outro problema sistêmico endereçado é a independência dos atributos mesmo quando as palavras são dependentes. *ComplementNB* impede que classes com mais dependências dominem, por meio da normalização dos pesos de classificação. Segundo SCIKIT-LEARN (2019), o *ComplementNB* é particularmente adequado para conjuntos de dados desbalanceados. Além disso, normalmente supera o *MultinomialNB* (geral-

mente por uma margem considerável) em tarefas de classificação de texto.

Ainda dentro do grupo de classificadores probabilísticos está o método da Regressão Logística. Apesar do nome, é um modelo linear para classificação. Neste modelo as probabilidades dos resultados de uma amostra são modeladas usando uma função logística. Na Scikit-learn o modelo é implementado na classe *LogisticRegression* que na sua configuração padrão aplica a estratégia "one-vs-rest" (um-contra-todos) na classificação *multiclass* .

2.5.4 Lineares

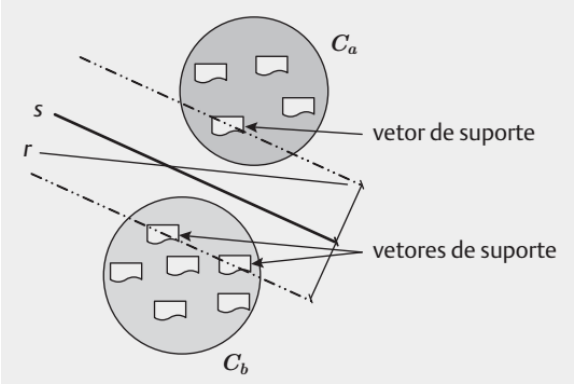
Modelos lineares constroem um hiperplano para representar uma separação linear entre classes. Em um espaço d-dimensional, um hiperplano é um subespaço de dimensão d-1. No plano bidimensional, um hiperplano é uma linha reta. No espaço tridimensional, um hiperplano é um plano regular.

O classificador Máquina de Vetores de Suporte (SVM, *Support Vector Machine*) (CORTES; VAPNIK, 1995) implementa essa ideia maximizando a margem definida pelo hiperplano de separação. No exemplo bidimensional da Figura 13, a linha s maximiza as distâncias aos documentos mais próximos nas classes C_a e C_b em comparação com a linha s e constitui o hiperplano de decisão usado para classificar novos documentos (BAEZA-YATES; RIBEIRO-NETO, 2013). As linhas paralelas pontilhadas delimitam a região onde devemos procurar por uma solução e são denominadas hiperplanos delimitadores. Documentos que pertencem a um hiperplano delimitador são chamados de vetores de suporte.

Para casos não linearmente separáveis, quando não há hiperplano que separe os pontos de dados em dois conjuntos disjuntos, pode-se projetar uma solução que permita ao classificador cometer alguns erros (abordagem de margem flexível) ou mapear os dados em um espaço dimensional mais alto no qual são linearmente separáveis (abordagem de *kernel*) (BAEZA-YATES; RIBEIRO-NETO, 2013).

Na Scikit-learn os modelos *SVC* e *LinearSVC* implementam SVM. O *SVC* usa a abordagem "um-contra-um" (*one-against-one*) para classificação *multiclass*: se existirem *n_ classes*, então *n_ classificadores* são construídos, de acordo com a equação 2.14, e cada um treina dados de duas classes.

Figura 13 – SVM bidimensional



Fonte: Extraído de Baeza-Yates e Ribeiro-Neto (2013).

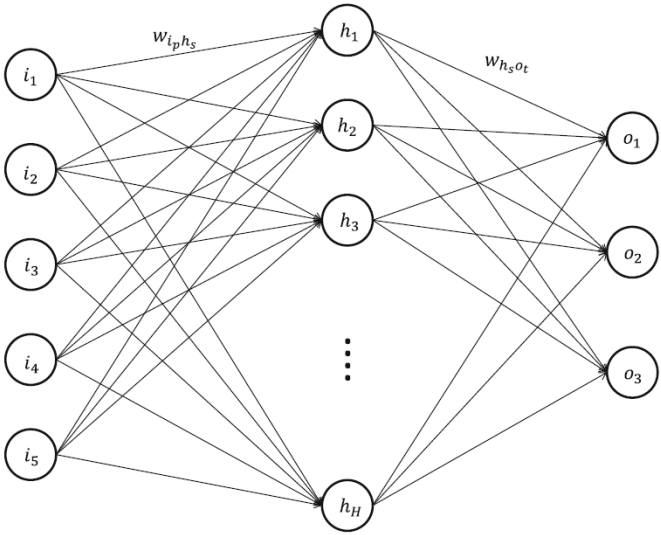
$$n\_classificadores = n\_classes * (n\_classes - 1)/2 \tag{2.14}$$

De outra forma, o *LinearSVC* adota a abordagem *multiclass* "um-contra-o-resto" (*one-vs-the-rest*): treina tantos modelos quantas classes existirem ($n\_classes = n\_classificadores$), entretanto se há duas classes apenas um modelo é treinado (SCIKIT-LEARN, 2019).

2.5.5 Redes Neurais Artificiais

Redes Neurais Artificiais são modelos lineares, mas são separados por conta da vasta família de classificadores. Métodos desse tipo são inspirados no funcionamento dos neurônios do cérebro humano. Um neurônio artificial é modelado como uma unidade que recebe um número de entradas ponderadas e fornece um resultado agregado. Essas unidades são colocadas juntas em uma rede que geralmente consiste em várias camadas: camadas de entrada e saída, bem como várias camadas ocultas (HERRERA et al., 2016). Dessa forma, modelos de Redes Neurais Artificiais podem modelar características de dados complexas. Um classificador de rede neural para instâncias em um espaço d-dimensional geralmente contém d nós de entrada (um para cada atributo) e m nós de saída, onde m é o número de classes possíveis. Um exemplo de rede é apresentado na Figura 14, onde cada nó está associado a um peso.

Figura 14 – Rede Neural Multicamadas



Fonte: Extraído de Vluymans (2019).

De acordo com Haykin (2009), a maneira como os neurônios de uma rede neural são estruturados está intimamente ligada ao algoritmo de aprendizado usado para treinar a rede. O autor identifica três arquiteturas de rede: Redes *Feedforward* de Camada Única, Redes *Feedforward* Multicamadas e Redes Recorrentes. Os tipos *feedforwards* só podem ter sinapses de saída com outros neurônios que estejam em camadas posteriores, já as redes recorrentes têm pelo menos um *loop* de *feedback* (retorno a neurônios de camadas anteriores).

Um perceptron multicamada (MLP, *Multi-layer Perceptron*) contém uma ou mais camadas ocultas. Segundo (SCIKIT-LEARN, 2019) o MLP pode ajustar um modelo não linear aos dados de treinamento. A biblioteca Scikit-learn implementa a classe *MLPClassifier* usando *Backpropagation* no treinamento.

2.5.6 Métodos de Ensemble

Os métodos de *ensemble* contemplam vários modelos de classificação a partir do mesmo conjunto de treinamento possibilitando a captura de diferentes comportamentos de classificação o que pode pro-

porcionar melhores resultados. Isso pode ser obtido usando diferentes algoritmos de aprendizado, fazendo várias versões modificadas do conjunto de treinamento ou ambos. Para classificar uma nova instância x, cada membro do conjunto calcula uma previsão e a atribuição da classe final é derivada por meio de um procedimento de agregação (por exemplo, um voto majoritário).

No SCIKIT-LEARN (2019) duas famílias de métodos *ensemble* são identificadas:

1. Métodos de média: constrói-se de maneira independente vários estimadores e, em seguida, calcula-se a média de suas previsões. Em média, o estimador combinado é geralmente melhor que qualquer um dos estimadores de base porque sua variância é reduzida.

2. Métodos de *boosting*: os estimadores de base são construídos sequencialmente e um deles tenta reduzir o viés do estimador combinado. A motivação é combinar vários modelos fracos para produzir um conjunto poderoso.

Com respeito ao grupo 1, a classe *RandomForestClassifier* da Scikit-learn, implementa um meta-estimador que se ajusta a vários classificadores de árvore de decisão em várias subamostras do conjunto de dados e usa a média para melhorar a precisão preditiva e controlar o ajuste excessivo (*overfitting*) (SCIKIT-LEARN, 2019).

No grupo 2, a biblioteca Sciki-learn disponibiliza a classe *AdaBoostClassifier* que implementa o algoritmo introduzido por Freund e Schapire (1997). O nome AdaBoost deriva de *Adaptive Boosting* (impulso ou estímulo adaptativo, em português). Segundo SCIKIT-LEARN (2019), o princípio central do *AdaBoost* é ajustar uma sequência de aprendizes fracos (por exemplo, pequenas árvores de decisão) em versões repetidamente modificadas dos dados. As previsões de todos eles são então combinadas por meio de um sistema de voto majoritário ponderado (ou soma) para produzir a previsão final. As modificações de dados em cada iteração de reforço consistem em aplicar pesos a cada uma das amostras de treinamento. Para cada iteração sucessiva, os pesos são modificados individualmente e o algoritmo de aprendizado é reaplicado aos dados ponderados novamente. Os exemplos de treinamento que foram incorretamente previstos na etapa anterior têm seus pesos aumentados, ao passo que para os que foram preditos corretamente os pesos são reduzidos. À medida que as iterações prosseguem, exemplos difíceis de prever recebem uma influência cada vez maior.

3 TRABALHOS RELACIONADOS

Existe um bom volume de pesquisas publicadas aplicando dados de RSO ao contexto das cidades. Em relação ao escopo e propósitos deste trabalho, é importante entender:

- como a mídia social tem sido usada para lidar com questões urbanas,
- se existem casos que integrem dados de RSO à serviços da cidade e
- se há casos que classifiquem dados das RSO dentro de modelos de indicadores para Cidades Inteligentes.

Nos três cenários, o interesse principal é pelo tipo de classificação realizada nessas pesquisas, procurando especialmente por abordagens que utilizam aprendizagem de máquina e técnicas de PLN.

3.1 RSO NAS QUESTÕES URBANAS

Existem várias pesquisas usando dados de mídia social para estudar riscos naturais, eventos epidêmicos ou de emergência (MIDDLETON; MIDDLETON; MODAFFERI, 2014; KIM; HASTAK, 2018), e o Twitter tem sido usado como fonte de dados em várias delas. Essas pesquisas geralmente são baseadas na detecção de eventos.

Em relação aos dados de RSO relacionados ao ambiente urbano, uma interessante pesquisa de detecção de eventos é apresentada por Nolasco e Oliveira (2019), onde os *tweets* são submetidos a algoritmos de classificação de tópico para representar eventos e subeventos com rótulos específicos. Essa abordagem fornece maneiras de entender melhor os eventos e seus componentes.

Existem muitas pesquisas relacionadas a Redes Sociais Baseadas em Localização (*Location-based Social Networks*, LBSN) (OLIVEIRA, 2016; DOMÍNGUEZ et al., 2017; CEREZO-COSTAS et al., 2018; LIU et al., 2019; SI; ZHANG; LIU, 2019) que se referem às RSO que incorporam informações de localização em conteúdos compartilhados (ROICK; HEUSER, 2013). A localização é uma informação importante do contexto do usuário e muito conhecimento pode ser aprendido com a mesma, no entanto, é comum que os usuários não compartilhem informações de localização. De acordo com Middleton, Middleton e Modafferi (2014),

apenas cerca de 1 % de todos os *tweets* contêm localização. No trabalho desenvolvido nesta dissertação, dos 7.021 *tweets* de cidadãos coletados, apenas um *tweet* continha coordenadas geográficas (latitude e longitude), e apenas cerca de 40 % deles foram postados de contas onde os usuários informaram a localização da cidade (Porto Alegre).

Métodos computacionais para fornecer identificação automatizada de questões urbanas relacionadas à infraestrutura das cidades foram investigados por Oliveira (2016) usando postagens de mídia social em inglês. Dois conjuntos de dados foram usados: um do Twitter com 1.494 *tweets* com *geotags* manualmente rotulados (OLIVEIRA et al., 2017) e um conjunto de dados com 18.090 mensagens do FixMyStreet (aplicativo onde pessoas de cidades europeias relatam problemas de Infraestrutura). A pesquisa foi concentrada em dados das cidades de Dublin e Londres. O autor propôs uma Ontologia de Domínio para Questões Urbanas (*Urban Issues Domain Ontology*, UIDO) para permitir a identificação e classificação de questões urbanas em uma abordagem automatizada com foco nas facetas temáticas e geográficas. Para a faceta geográfica, um *geoparser* foi usado para atribuir latitude e longitude aos *tweets*. O desempenho temático da ontologia foi comparado com as abordagens de aprendizagem de máquina: *Naïve Bayes*, *Naïve Bayes Multinomial*, *Random Forests* e *Support Vector Machines* (SVM).

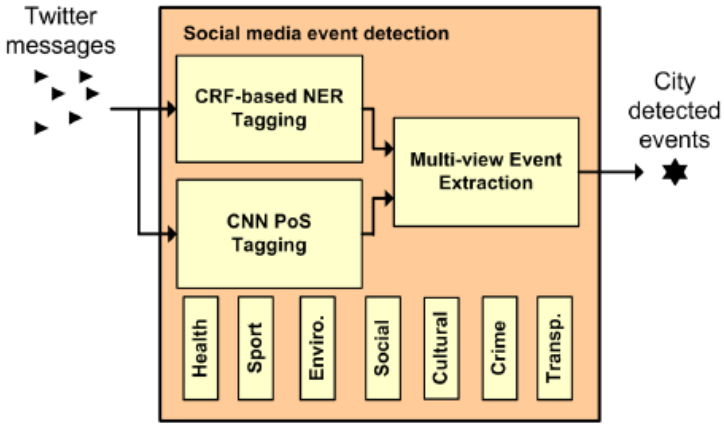
3.2 RSO INTEGRADAS AOS SERVIÇOS DA CIDADE

A "inteligência" de uma cidade está relacionada com a capacidade de reunir seus recursos para cumprir de forma efetiva e transparente as metas estabelecidas por meio da aplicação de uma nova geração de tecnologias da informação capazes de facilitar o planejamento, implementação e gerenciamento de serviços municipais (ISO, 2014). Exemplos de tais tecnologias são: Internet das Coisas (IoT, *Internet of Things*), computação em nuvem, *big data* e sistemas para a integração de informações geográficas e espaciais. Nessa direção, vários *frameworks* e plataformas já existem e disponibilizam dados de cidades que podem ser usados na implementação de serviços municipais por meio da análise de dados, como, por exemplo: CityPulse (PUIU et al., 2016), InterSCity (ESPOSTE et al., 2019), IES Cities (AGUILERA et al., 2017), RADICAL, SCDAP e outros (OSMAN, 2019).

Em especial a plataforma CityPulse além de processar informações de diversos sensores, contém uma unidade dedicada a mensagens

com *geotag* da rede social Twitter representada na Figura 15. Essa unidade tem 3 componentes. O *Conditional Random Field* (CRF) com *Name Entity Recognition* (NER) reconhece entidades nomeadas e atribui tags de categorias de eventos usando palavras chaves. Os tipos de eventos contemplados são Saúde, Esporte, Meio Ambiente, Social, Cultural, Crime, Transportes. O componente *Convolutional Neural Network* (CNN) gera as *tags* de partes do discurso (POS, *Part-Of-Speech*) usadas para identificar a classe gramatical das palavras. O terceiro componente *Multi-view Event Extraction* valida as *tags* oriundas dos dois componentes anteriores e resolve ambiguidades para os casos de *tweets* pertencendo a mais de um tipo de evento. A plataforma CityPulse anuncia que pode ser usada em qualquer linguagem visto que traduz os *tweets* para inglês usando a API *Google-translate*. Os autores também mencionam que a plataforma contém um módulo de agregação dos dados onde efetuam o *clustering*.

Figura 15 – Plataforma CityPulse com dados do Twitter



Fonte: Extraído de Puiu et al. (2016).

As plataformas para Cidades Inteligentes normalmente implementam um conjunto de requisitos funcionais comuns na forma de serviços reutilizáveis para desenvolvimento de aplicativos, incluindo serviços para a integração de dispositivos da IoT, bem como para armazenamento e processamento de dados e para reconhecimento de contexto (ESPOSTE et al., 2019). As plataformas para coleta de dados de RSO devem prestar atenção nas alterações relacionadas à privacidade do

usuário e às limitações de extração de dados das diferentes RSO.

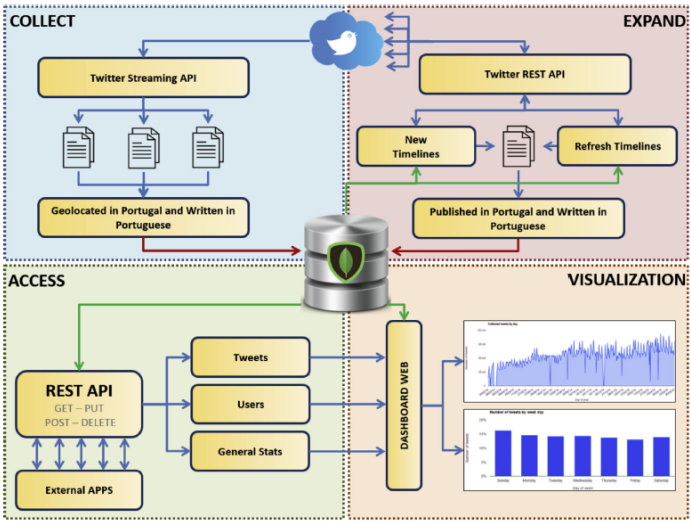
Por meio dos recursos destas plataformas, é possível permitir a implementação de serviços que consomem, analisam ou facilitam a visualização dos dados coletados. Esposte et al. (2019) observa que as arquiteturas das plataformas para Cidades Inteligentes podem atender a múltiplos requisitos não-funcionais, como adaptabilidade, privacidade, interoperabilidade e capacidade de evolução, de acordo com a especificidade do problema que pretendem resolver.

A plataforma *Intelligent Mining of Public Social Networks* (MISNIS, em português Mineração Inteligente de Redes Sociais Públicas) (CARVALHO et al., 2017), apresentada na Figura 16, é focada em dados de RSO e aborda coleta, armazenamento, gerenciamento, mineração e visualização de postagens do Twitter. Funciona 24 horas por dia e 7 dias por semana, usando a API de *Streaming* do Twitter para coletar *tweets* geolocalizados em Portugal, com o objetivo de facilitar um usuário não técnico a extrair um determinado tópico de um corpus de *tweets* muito grande. A plataforma aborda três tipos diferentes de análise: Influência do usuário, Análise de sentimentos e Detecção de Tópicos. As funções de classificação de texto utilizam o que o autor denomina *Fuzzy Fingerprints* baseada em lógica fuzzy, onde um conjunto de textos associados a uma dada classe é usado para construir a impressão digital da classe e funções de similaridade são aplicadas para enquadrar a impressão digital de novos documentos a uma destas classes.

As pesquisas de Cidades Inteligentes, que propõem serviços usando dados das RSO, geralmente se concentram em alguns domínios específicos. Há muitos trabalhos relacionados à dimensão Transportes, que inclui Mobilidade, como pode ser verificado em Dabiri e Heaslip (2018), Panagiotou et al. (2016), Domínguez et al. (2017), Kousiouris et al. (2018). Em Kousiouris et al. (2018) há uma proposta para integrar a plataforma Smart City COSMOS, oriunda de um projeto europeu patrocinado por um consórcio entre várias empresas e universidades, com RSO e fornecer um serviço para identificação de eventos definidos como *Large Crowd Concentration* (LCC, Grande Concentração de Multidões) em uma determinada área. A ideia é desenvolver uma experiência personalizada para que o cidadão possa identificar esses LCC e tomar decisões relacionadas às suas metas de mobilidade, em especial para transporte de pessoas com limitações de cognição. A identificação é baseada na observação dos picos de atividade do Twitter em comparação com os dados históricos e local de interesse.

Outro exemplo ligado ao gerenciamento de dados urbanos da

Figura 16 – Plataforma MISNIS com dados do Twitter

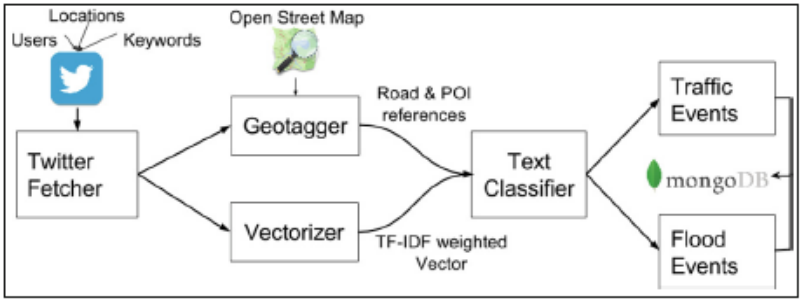


Fonte: Extraído de Carvalho et al. (2017).

cidade de Dublin pode ser encontrado em Panagiotou et al. (2016). Na Figura 17 está representado o componente de análise dos dados do Twitter, onde foi implementado um *geotagger* para atribuir latitude e longitude aos *tweets* de acordo com informações geográficas referenciadas no *post*, como nomes de estradas e *Points of Interest* (POI, Pontos de Interesse). Um classificador SVM foi usado nos *tweets* que puderam ser georreferenciados pelo *geotagger* O Conselho Municipal de Dublin (DCC, *Dublin City Council*) opera um centro de gerenciamento de tráfego analisando informações de várias fontes. Os autores implementaram soluções para detecção de eventos e resposta a emergências usando dados do Twitter. Um exemplo é o aplicativo *CrowdAlert*, o qual permite que os cidadãos e os operadores do DCC observem os eventos em andamento identificados a partir do sistema em tempo real.

A opinião pública sobre os serviços da cidade também pode ser rastreada a partir das RSO. A dimensão Saúde da cidade atrai a atenção de diversas pesquisas (BELLO-ORGAZ; HERNANDEZ-CASTRO; CAMACHO, 2017; D'ANDREA et al., 2018). Uma classificação foi realizada no Twitter em D'Andrea et al. (2018) com o objetivo de monitorar a opinião pública sobre a vacinação na Itália. A abordagem proposta processa *tweets* relacionados à vacinação e emprega um modelo SVM

Figura 17 – Componente de Análise do Twitter



Fonte: Extraído de Panagiotou et al. (2016).

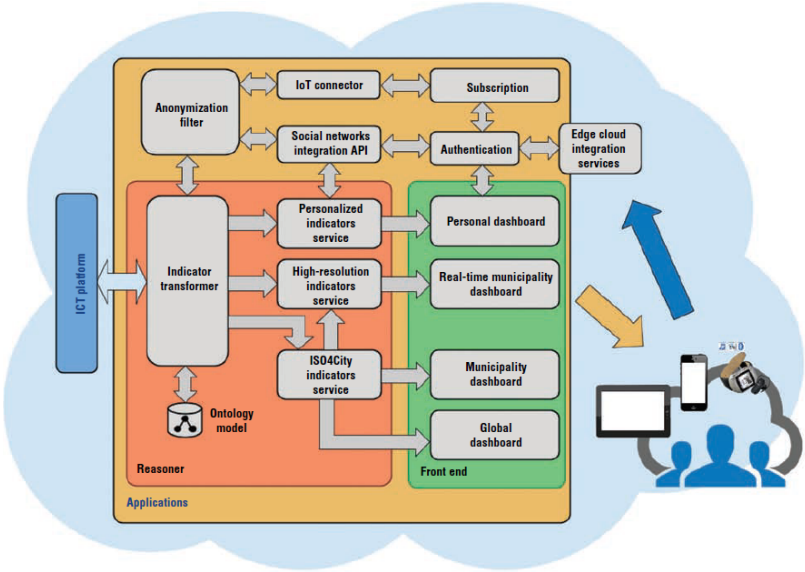
para classificar os mesmos como pertencentes a uma das três classes: a favor da vacinação, contra e neutra.

3.3 RSO E MODELOS DE CIDADES INTELIGENTES

A ideia de uma plataforma inspirada nos indicadores da ISO 37120 e conectada a várias fontes de dados da cidade é apresentada em Zdraveski et al. (2017). Os autores qualificam os indicadores da ISO como de "baixa resolução" porque são estáticos ou mudam lentamente (anual, trimestral ou mensalmente). Eles propõem o desenvolvimento de indicadores *proxy* (intermediários) conectados a fontes dinâmicas de dados da cidade, como sensores, redes sociais, notícias, blogs e fontes municipais de sistemas de informação. Os indicadores substitutos melhorariam a resolução dos indicadores ISO. Os autores mencionam uma API de "integração de redes sociais" que conectaria cidadãos a contas de redes sociais e classificaria os usuários com base em seu envolvimento na melhoria dos serviços e qualidade de vida da cidade. Os autores igualmente citam uma ontologia dentro de um raciocinador responsável pela transformação dos dados em indicadores, que pode ser visualizada na Figura 18. Entretanto, eles não descrevem como essas tarefas aconteceriam e quais RSO seriam consideradas. O trabalho está mais focado em propor uma arquitetura inspirada no modelo ISO.

A tarefa de mineração de políticas urbanas é apresentada por Puri et al. (2018) e pode ser visualizada na Figura 19. Os autores analisam as ordenações (regulamentos, decretos, portarias ou leis locais) oriundas de websites com relação à reação da opinião pública sobre as

Figura 18 – Arquitetura de Plataforma - Smart City e ISO 37120



Fonte: Extraído de Zdraveski et al. (2017).

mesmas expressa no Twitter. Para fazer essa conexão, as ordenações e os *tweets* foram categorizados em *Smart City Characteristics* (SCC, Características da Cidade Inteligente), como *Smart Mobility* e *Smart Living*. Essas categorias são dimensões do modelo europeu de indicadores para Cidades Inteligentes (TUWIEN, 2015).

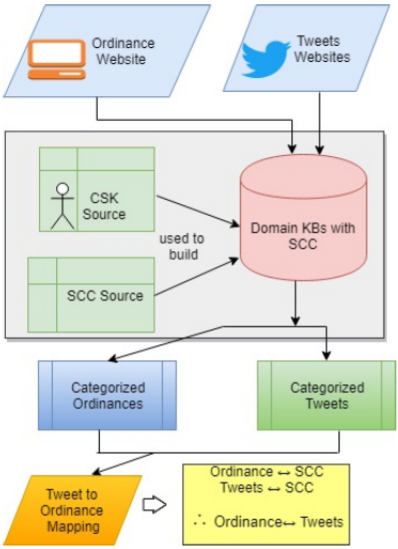
A abordagem empregada foi a construção de bases de conhecimento específicas por domínio, contendo os termos relacionados. Essas bases foram construídas, usando repositórios de conhecimento de senso comum (*Common Sense Knowledge*, CSK), como o WebChild[1] e o WordNet[2].

3.4 CONSIDERAÇÕES DO CAPÍTULO

Existem diversos trabalhos a respeito da aplicação das RSO no ecossistema das cidades, em especial com dados do Twitter. Nas Tabelas 1 e 2 é possível verificar um resumo dos trabalhos avaliados e as

[1] https://gate.d5.mpi-inf.mpg.de/webchild/
[2] https://wordnet.princeton.edu/

Figura 19 – Proposta para classificação usando CSK



Fonte: Extraído de Puri et al. (2018).

respectivas dimensões e estratégias de análise aplicadas.

Os trabalhos de Puri et al. (2018) e Oliveira (2016) tem objetivos próximos dos buscados na pesquisa que embasa esta dissertação. Todavia, o trabalho apresentado por Puri et al. (2018) não está explorando abordagens de aprendizagem de máquina, ele usa o modelo europeu com seis classes e não considera outros modelos de Cidades Inteligentes.

A classificação das facetas temáticas desenvolvida por Oliveira (2016) está muito relacionada a presente pesquisa e serve de inspiração. No entanto, o autor foca principalmente na abordagem por ontologia e não investiga profundamente as opções de aprendizagem de máquina. Além disso, o trabalho está concentrado apenas em questões de Infraestrutura, usando somente seis categorias de classificação para dados do FixMyStreet, e para a base do Twitter foi efetuada classificação binária (considera se o *tweet* tem ou não questões urbanas de Infraestrutura, sem classificar nas categorias).

Tabela 1 – Principais Trabalhos Relacionados (continua)

Artigo	Tipo	Goal	Abordagem	Temas	RSO	PU	LT	IS	SCM	PLT
Middleton et. al (2014)	ED	Plataforma de crise efetuando geoparsing em tempo real	Análises estatísticas offline para ED	Resposta a desastres	Twitter	S	N	S	N	S
Nolasco e Oliveira (2019)	ED	Detectar subeventos e assim entender melhor os eventos e seus componentes	Aprendizagem de máquina não supervisionada com LDA (Latent Dirichlet Allocation)	Saúde e Protestos	Twitter	S	S	N	N	N
Oliveira (2016)	TC	classificar as mensagens que se referem à questões de infraestrutura urbana e identificar localização das mesmas com geoparser.	Proposta de Ontologia urbana para classificar mensagens comparada à classificação supervisionada (SVM e outros)	Infraestrutura	Twitter	S	S	N	N	N
Cerezo-Costas et al. (2018)	ED	(i) a detecção de comportamento inesperado na cidade e (ii) a análise dos posts para inferir o que está acontecendo	(i) Clustering para detectar comportamentos inesperados nas postagens (*Crowd detection*) e (ii) técnicas de agregação de conteúdo para o segundo goal	Mobilidade	Instagram	S	S	S	N	N
Puiu et al. (2016)	ED TC	Apresentar o framework CityPulse, descrever seus componentes, demonstrar exemplos de aplicações	Usam técnicas de NLP para classificar os *tweets* (NER, POS) e clustering para agregação dos dados	Crimes, Saúde, Esportes, Meio Ambiente, Cultura, Transportes	Twitter	S	S	S	N	S
Carvalho et al. (2017)	TC TD	Plataforma MISNIS, que disponibiliza *tweets* e facilita um usuário não técnico a extrair determinado tópico do corpus e verificar análises	Lógica fuzzy (*fuzzy fingerprint* para classificar) e funções de similaridade e NLP	na	Twitter	N	S	S	N	S

Tipo: ED-Event Detection, TP-Topic Detection, TC-Topic Classification
PU: identifica Problemas Urbanos
LT: depende da Localização do *Tweet*/usuário
IS: prevê Integração à Serviços da cidade
SCM: faz referência a um modelo de Smart City
PLT: propõe ou implementa uma plataforma

Fonte: Elaborado pela autora.

Tabela 2 – Principais Trabalhos Relacionados (conclusão)

Artigo	Tipo	Goal	Abordagem	Temas	RSO	PU	LT	IS	SCM	PLT
Panagiotou et al. (2016)	ED	Desenvolver um conjunto de técnicas para gerenciamento de dados urbanos em um cenário real na cidade de Dublin	Somente analisa *tweets* onde é capaz de utilizar o *geotagger* com sucesso. Utiliza SVM na identificação de eventos de trânsito e enchentes. TF-IDF para ponderação dos atributos.	Transporte	Twitter	S	N	S	N	N
Kousiouris et al. (2018)	ED	Identificar concentrações de pessoas para facilitar mobilidade de pessoas com restrições de mobilidade ou cognitivas	Menciona aplicação de aprendizagem de máquina usando módulos do COSMOS e aplicando Complex Event Processing (CEP)	Transporte	Twitter	N	S	S	N	S
D'Andrea et al. (2018)	TC	Sistema para identificar tendências na opinião pública sobre vacinação	SVM para classificar análise de sentimento	Saúde	Twitter	S	S	S	N	N
Zdraveski et al. (2017)	NA	Proposta de uma plataforma, orientada pela norma ISO 37120, que tem como objetivo adquirir e processar dados de fontes e sensores heterogêneos, implementados na nuvem	Foco na arquitetura, menciona um raciocinador que faria uso de uma ontologia capaz de transformar os dados nos indicadores da ISO	NA	NA	N	N	N	S	S
Puri et al. (2018)	TC	Classificar os *tweets* dentro de categorias de *Smart Cities* do modelo europeu para conectar as ordenações/políticas relacionadas àquela categoria	Aplicam técnicas de PLN para limpar as mensagens. Usam bases CSK para extrair termos para as 6 dimensões do modelo europeu e efetuam o mapeamento dos *tweets*	Todos do modelo europeu	Twiter	S	S	N	S	N

Tipo: ED-Event Detection, TP-Topic Detection, TC-Topic Classification
PU: identifica Problemas Urbanos
LT: depende da Localização do *Tweet*/usuário
IS: prevê Integração à Serviços da cidade
SCM: faz referência a um modelo de Smart City
PLT: propõe ou implementa uma plataforma

Fonte: Elaborado pela autora.

4 MODELO PROPOSTO

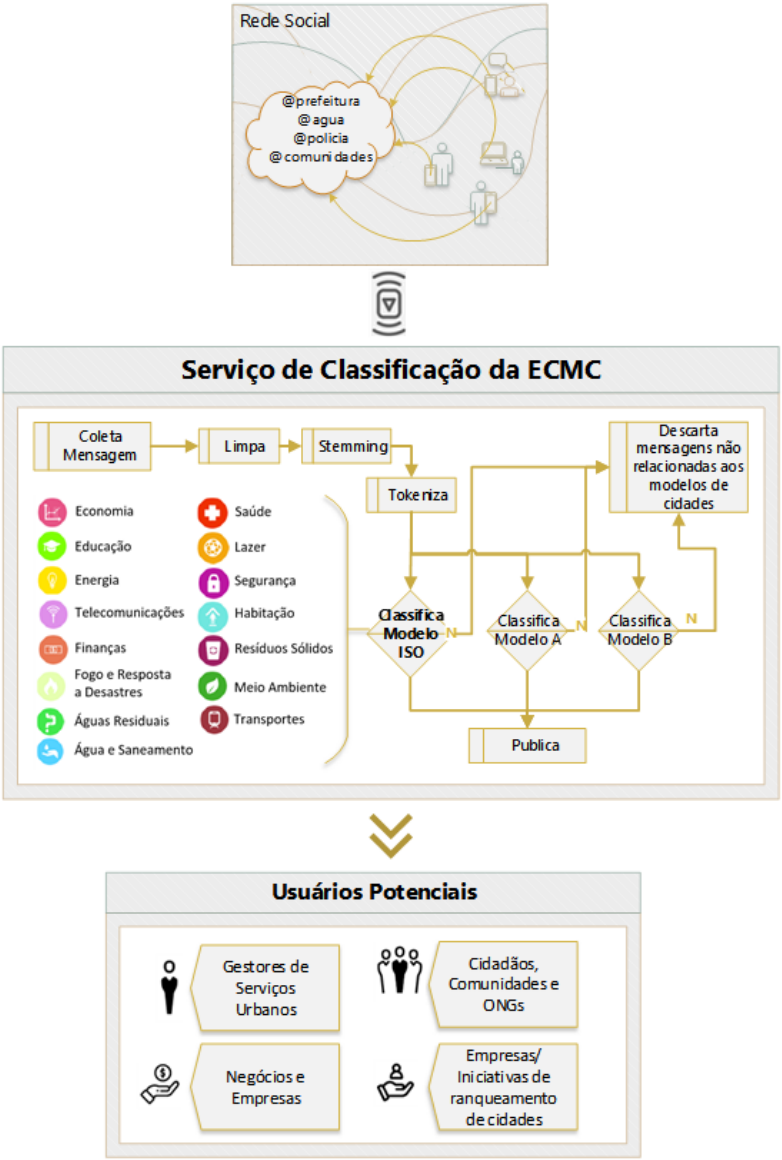
Este capítulo apresenta propostas capazes de responder às perguntas desta pesquisa e atingir os objetivos estabelecidos. São detalhados a proposta de Estrutura de Classificação de Mensagens das redes sociais nas dimensões da Cidade (ECMC), os processos e experimentos de aprendizagem de máquina executados a fim de determinar o Modelo de Classificação Supervisionada (MCS) para ISO 37120 e a heurística empregada na análise dos desafios em incorporar outros modelos de indicadores de Cidades Inteligentes ao ECMC.

4.1 PROPOSTA DE ECMC

Na Figura 20 é possível verificar que a ECMC é baseada em três componentes. O primeiro se refere à estratégia usada para coletar as mensagens relevantes dos cidadãos nas RSO sobre as dimensões da cidade. São usadas contas oficiais ou amplamente reconhecidas pelos cidadãos da localidade em questão e que tenham conexão com as dimensões do modelo de Cidades Inteligentes a ser utilizado. Todas as mensagens enviadas para estas contas são passíveis de classificação. Depois de postada mensagem mencionando pelo menos uma das contas de interesse, ela é enviada ao segundo componente, o serviço de classificação, onde é pré-processada e transformada em um vetor de tokens. Os classificadores dos modelos de cidades são aplicados sobre o vetor, determinando se ele se enquadra a uma das dimensões dos modelos. Caso nenhuma dimensão da cidade seja encontrada, a mensagem será descartada, do contrário, ela será encaminhada para publicação. O terceiro componente representa os consumidores dos dados publicados, que abrange especialmente gestores dos serviços públicos mas também empresas, comunidades, cidadãos, a própria ISO e outras empresas que efetuam ranking de cidades ou prestam serviços para as mesmas.

É visível a importância do segundo componente, onde está embutida a inteligência de classificação. Para cada modelo de Cidade Inteligente disponível no serviço de classificação é necessário definir o melhor Modelo de Classificação Supervisionado (MCS), treiná-lo e testá-lo para que possa categorizar as mensagens dos cidadãos oriundas das RSO dentro das dimensões do respectivo modelo. Este trabalho foca na ISO 37120 como o modelo de Cidade Inteligente a ser explorado e nas RSO Twitter e Colab que serão as fontes das mensagens.

Figura 20 – Estrutura de Classificação de Mensagens da Cidade



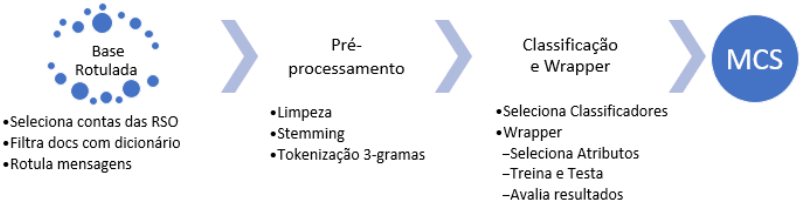
Fonte: Elaborado pela autora.

4.2 METODOLOGIA PARA DETERMINAR MCS

O processo para determinar o MCS está resumido na Figura 21 e cada passo será detalhado nas próximas subseções. A Subseção 4.2.1 detalha o desenvolvimento da "Base Rotulada", descrevendo os filtros aplicados nas RSOs utilizadas neste trabalho, bem como o processo de classificação manual das mensagens coletadas e geração do dicionário de termos relacionados a ISO 37120. O "Pré-processamento" é descrito na Subseção 4.2.2 detalhando as técnicas utilizadas no tratamento das mensagens da base rotulada como limpeza, padronização do texto, filtragem, stemming e tokenização.

A abordagem de classificação é baseada em uma estratégia de *wrapper*, que combina dois parâmetros: os algoritmos e a técnica de seleção de atributos (estas técnicas são selecionadas de forma a otimizar os classificadores). Assim, os processos descritos em "Classificação e Wrapper" se referem primeiramente a seleção dos algoritmos dentre vários disponíveis na biblioteca Scikit-learn, descrito na Subseção 4.2.3. A seguir, na Subseção 4.2.4, os algoritmos com melhores resultados na escolha preliminar são combinados as técnicas para seleção dos atributos de forma a obter as configurações que geram os melhores resultados e assim determinar o melhor MCS.

Figura 21 – Processo de determinação do MCS



Fonte: Elaborado pela autora.

Dois algoritmos pré-selecionados foram aplicados sobre os dados e combinados a cinco técnicas para extrair e selecionar os atributos que efetivamente contribuem para melhorar os resultados da classificação. A decisão final do melhor MCS é baseada primeiramente nos resultados da F1-macro, e depois na F1-micro e no desempenho das categorias internas.

4.2.1 Coleta de dados

As postagens compartilhadas pelos cidadãos nas RSO não estão restritas a problemas da cidade. A seleção das RSO levou em consideração sua capacidade de incentivar publicações sobre o ambiente da cidade. É importante que a rede social permita identificar a cidade à qual o *post* se refere para poder medir a voz da população local dentro das dimensões da ISO.

4.2.1.1 Dados do Twitter

Como a porcentagem de *tweets* com *geotags* é muito baixa, optou-se por usar apenas *tweets* que mencionam algumas contas selecionadas relacionadas aos serviços da cidade. A cidade de Porto Alegre foi usada como estudo de caso para o Twitter pois o governo local tem um perfil oficial ativo e existem várias contas no Twitter relacionadas à serviços municipais, como água, trânsito, lixo, defesa civil, etc. Muitos usuários mencionam esses perfis em suas postagens, procurando uma maneira de direcionar sua mensagem para uma conta do Twitter que é interpretada pelo usuário como responsável pelo tópico ou que pode contribuir para o encaminhamento do problema. Na Figura 22 estão alguns exemplos de *tweets* que foram rotulados dentro das categorias ISO.

O Twitter possui *Application Programming Interfaces* (APIs) públicas que permitem acesso às postagens feitas por contas públicas (contas que não mantém restrição de acesso às suas postagens). Neste trabalho foi empregada a biblioteca Tweepy[1] que consome as APIs do Twitter. Utilizou-se a API de *streaming* do Twitter para fazer a extração de mensagens em tempo real. A API de *streaming* é diferente da API REST porque esta última é usada para extrair dados passados do Twitter, enquanto que a API de *streaming* obtém as mensagens através de uma sessão persistente e é útil para obter um alto volume de *tweets* ou para criar um fluxo ativo de mensagens. Isso permite que a API de *streaming* baixe mais dados em tempo real do que poderia ser feito usando a API REST (TWEEPY, 2019).

O script de extração de dados do Twitter foi desenvolvido em Python[2] versão 3.6. As macro etapas do script podem ser visualizadas na Figura 23. Não foram utilizadas palavras chaves como filtro, apenas restringiu-se a coleta a um conjunto de contas previamente seleciona-

[1] http://www.tweepy.org/
[2] https://www.python.org/

Figura 22 – Exemplos de Tweets dos Cidadãos

ISO Dimension	Tweet
Água e Saneamento	@dmaepoa Existe uma possibilidade de assinatura de avisos acerca de falta de água disponível nos serviços de mídia social?
Águas Residuais	@dmaepoa alerta de esgoto transbordando em frente a Ceee Joaquim Vilanova! Bom Domingo a todos!
Economia	@portoalegre24h Mais do que justo, pois o preço do diesel recuou, e as empresas, com a conivência da EPTC, prefeitura e órgãos trabalhistas, diminuíram o número de horários das linhas, causando desemprego, inclusive.
Educação	@portoalegre24h Várias pessoas não conseguem se inscrever o site tá dando erro e o MEC não se pronuncia
Energia	@Prefeitura_POA alguma explicação sobre a falta de luz geral? Estamos no bairro Auxiliadora
Finanças	@Prefeitura_POA O IPTU que eu pago é muito maior do que o retorno que recebo em cuidados com a rua..bairro..cidade! Abandonada!! E ainda dizem estar defasado??😠
Habitação	@curtaramiro Para onde foram realocados os moradores de rua que ali estavam!? Tenho certeza que o foram para um projeto de ressocialização! Ou não secretário!?
Lazer	@smcprefpoa Fale-me mais sobre a EXTINÇÃO do fundo para restauro de prédios históricos Lamento da mesma forma
Meio Ambiente	@curtaramiro Vocês tem é que plantar mais árvores. Está feio. Uma clareira aberta.Cada vez menos árvores,
Resíduos Sólidos	@DMLU_POA Olhem este lixo quase caindo está lá a um ano, amanhã vou lá fotografar novamente. RECOLHAM O LIXO DAS MARGENS DO ARROIO DILÚVIO!
Saúde	@saudepoa Bairro serraria Vila dos sargentos sem posto de saúde a mais de um ano queremos providências
Segurança	@portoalegre24h nada sobre o tiroteio na rua em frente a puc? vi gente morta dentro de um carro ali hj pela manha!
Fogo e Resposta a Desastres	@portoalegre24h Aqui na sona Norte foi um ciclone com fortes ventos desgalhando prédios e casas e o vento continua bairro Rubem Berta e nova gleba
Transportes	@EPTC_POA Uma bela homenagem ao pedestre seria consertar o telhado do terminal do triângulo na Assis Brasil e também uma cobertura para a parada da Carlos Gomes com a Anita.
Telecomunicações	@Prefeitura_POA Onde estão os apps prometidos na campanha ?

Fonte: Elaborado pela autora.

Figura 23 – Script para extração dos Tweets

```
from tweepy import Stream
from tweepy import OAuthHandler
from tweepy.streaming import StreamListener

auth = OAuthHandler(ckey, csecret)
auth.set_access_token(atoken, asecret)
twitterStream = Stream(auth, listener())
twitterStream.filter(follow=CONTAS)
```

Fonte: Elaborado pela autora.

das, as quais enviam e recebem cotidianamente mensagens relacionadas ao ambiente urbano ou à serviços públicos. A classe *listener* é uma classe herdada da classe *StreamListerner* do Tweepy. Na classe *listener* estão implementados a coleta e a persistência de dados. A API de *Streaming* do Twitter disponibiliza os dados no formato *JavaScript Object Notation* (JSON). Na Figura 24 pode ser visualizado um exemplo da estrutura de um *tweet* coletado.

Figura 24 – Exemplos de tweet no formato JSON


```
"created_at": "Sat Sep 08 14:14:04 +0000 2018",
"id": 1038430271417339900,
"id_str": "1038430271417339907",
"text": "Encerrado içamento do vão móvel da Ponte sobre o Guaíba (km
  97 da BR290)Sem previsão de novos içamentos para hoje (PRF191RS)",
"source": "<a href=\"http://placeholder.com\" rel=\"nofollow\"
  >Informações de Trânsito</a>",
"truncated": false,
"in_reply_to_status_id": null,
"in_reply_to_status_id_str": null,
"in_reply_to_user_id": null,
"in_reply_to_user_id_str": null,
"in_reply_to_screen_name": null,
"user": {⌨},
"geo": null,
"coordinates": null,
"place": null,
"contributors": null,
"is_quote_status": false,
"quote_count": 0,
"reply_count": 0,
"retweet_count": 0,
"favorite_count": 0,
"entities": {⟺},
"favorited": false,
"retweeted": false,
"filter_level": "low",
"lang": "pt",
"timestamp_ms": "1536416044464"
```

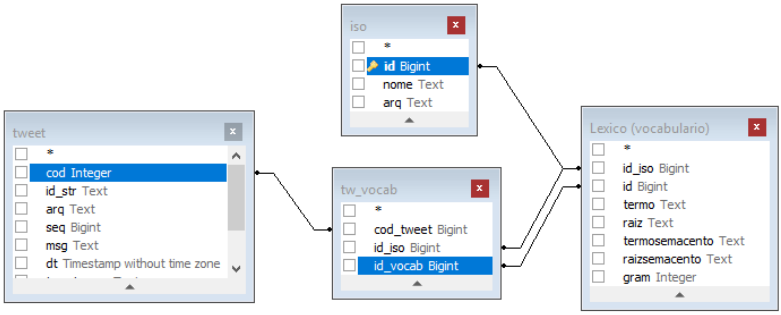

Fonte: Elaborado pela autora.

Todos os arquivos JSON coletados foram carregados em um banco de dados Postgres[3] versão 10 para facilitar algumas análises. Parte do esquema deste banco de dados pode ser visualizado na Figura 25. O processo de coleta foi executado durante os últimos seis meses de 2018, com mais de 50 mil *tweets* recuperados. Cerca de 7.000 *tweets*

[3] https://www.postgresql.org/

mencionaram os perfis selecionados e foram denominados como “tweets de cidadãos”. *Re-tweets* não foram considerados nem *tweets* provenientes das contas seguidas.

Figura 25 – Exemplo das tabelas do banco de dados



Fonte: Elaborado pela autora.

Um dicionário[4] de termos-chave da ISO foi construído manualmente para todas as categorias, usando a documentação disponível do padrão ISO 37120 e fazendo uma correlação manual com a documentação de outros modelos e de fontes: UN-Habitat[5], IBGE[6], PCS[7] e modelo europeu[8]. Foram avaliadas expressões e termos que correspondessem a cada dimensão. Na Figura 26 estão alguns exemplos.

Cada termo do dicionário tem até três palavras, desconsiderando-se artigos, preposições, conjunções, etc. Cada palavra foi reduzida a sua raiz usando o *stemmer* NLTK-Snowball[9] e, ao final, os acentos foram removidos. A redução total foi de cerca de 36% do número de termos resultando em 756 *stems* e sua distribuição pode ser vista na Figura 27.

A ISO 37120 foi criada em 2014 com 17 categorias, mas foram utilizadas apenas 15, não considerando Governança e Planejamento Urbano. A categoria Planejamento Urbano sobrepõe questões abordadas em outras dimensões. A dimensão Governança concentra-se no suporte para dados abertos e na disponibilidade e tempo de resposta do sis-

[4] https://github.com/lbencke/SC_ISO37120
[5] https://unhabitat.org
[6] https://sidra.ibge.gov.br/pesquisa/ids/tabelas
[7] https://www.cidadessustentaveis.org.br/
[8] http://www.smart-cities.eu/
[9] https://www.nltk.org/_modules/nltk/stem/snowball.html

Figura 26 – Exemplos de termos do dicionário para ISO

Meio Ambiente	**Saúde**	**Economia**
área de conservação	dengue	artesanato
área verde	desnutridos	custo de vida
muito barulho	doenças	desemprego
córrego	expectativa de vida	indústria
desmatamento	falta de leito	logística
qualidade do ar	medicamentos	preço da gasolina
sustentabilidade	mosquitos	sub-emprego
Educação	**Energia**	**Habitação**
analfabetos	elétrica	morador de rua
bibliotecas	eólica	área invadida
creches	fiação irregular	prédio ocupado
escolaridade	alta tensão	imóvel abandonado
matriculado	iluminação pública	população de rua
científico	sem luz	despejados
universitário	lâmpada	remoção das pessoas

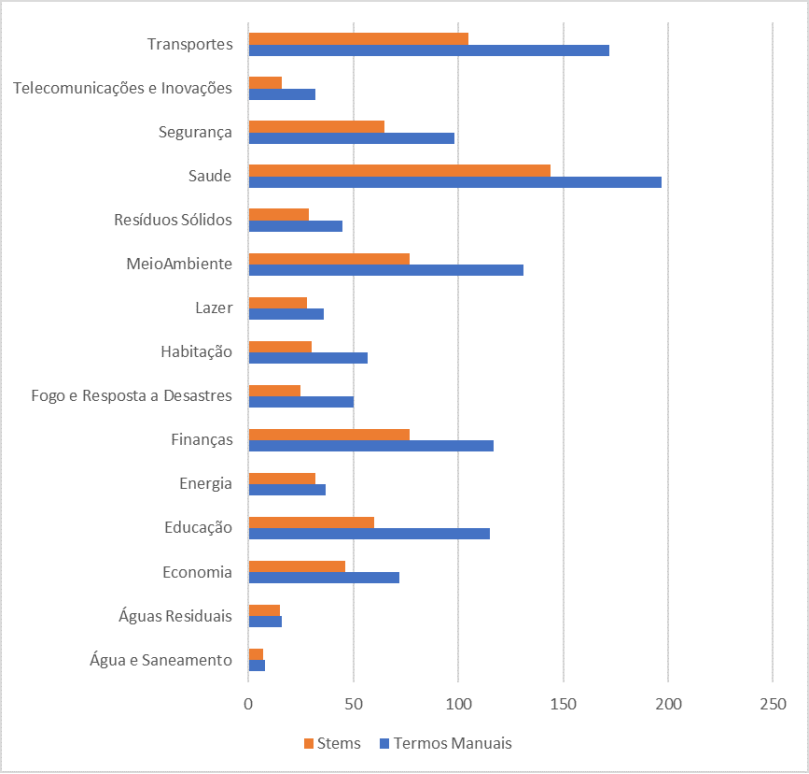
Fonte: Elaborado pela autora.

tema de reclamações. Este trabalho busca exatamente por este tipo de reclamação sobre os serviços da cidade, portanto, não se pode relacionar Governança a um serviço específico da cidade, mas sim a como gerenciar os dados e reclamações dos serviços de maneira geral. Por esse motivo, essa dimensão não é considerada pelos classificadores.

A base de dados rotulada do Twitter foi criada selecionando dois grupos de *tweets* de cidadãos: 1.002 *tweets* contendo pelo menos um dos termos do dicionário e 1.000 adicionais sem tal ocorrência. Apenas mensagens com mais de 25 caracteres (após o pré-processamento) foram consideradas. A maioria dos *tweets* que obtiveram alguma classificação foi vinculada a somente uma dimensão da ISO, enquanto que *tweets multilabel* correspondem a cerca de 2,5% da amostra e foram descartados restando 1950 *tweets*.

Como pode ser observado na Tabela 3, as classes estão altamente desbalanceadas dentro da base rotulada do Twitter, ou seja, há muita variação no número de amostras de cada categoria. Desconsiderando-se a classe Nenhum, as classes Transportes e Água e Saneamento têm o maior volume de *tweets* no conjunto de dados. A Classe Nenhum foi utilizada com o objetivo de identificar as mensagens que não pertencem a nenhuma classe ISO.

Figura 27 – Stemming sobre o Dicionário



Fonte: Elaborado pela autora.

Tabela 3 – Distribuição dos dados ao longo das dimensões ISO

Id Classe	**Dimensão**	**Bases**	
		Twitter	**Colab**
0	Água e Saneamento	8.1%	3.1%
1	Transportes	25.2%	47.0%
2	Energia	5.4%	13.1%
3	NENHUM*	49.5%	na
4	Lazer	1.0%	0.3%
5	Resíduos Sólidos	1.3%	13.1%
6	Economia	0.8%	0.4%
7	Finanças	0.9%	na
8	Águas Residuais	0.7%	5.8%
9	Segurança	2.2%	0.9%
10	Educação	0.6%	na
11	Meio Ambiente	5.4%	13.1%
12	Saúde	1.2%	3.0%
13	Telecomunicações	0.4%	na
14	Habitação	0.8%	3.5%
15	Fogo e Resposta à Desastres	1.1%	na
		100%	100%

***A classe NENHUM não pertence ao modelo ISO e corresponde a mensagens do Twitter que não se enquadram em nenhuma dimensão ISO.
Fonte: Elaborado pela autora.

4.2.1.2 Dados do Colab

Os conjunto de dados Colab foi recebido em arquivo texto, sem os dados de localização ou nome de usuário, no entanto, esses dados são públicos e aparecem em cada postagem no site. A rede Colab é muito focada em questões urbanas, portanto, representa um bom conjunto para comparar com o Twitter.

Quando os usuários compartilham uma postagem no Colab, eles classificam a postagem em uma das mais de 80 combinações de categorias e subcategorias internas. Na Tabela 4 encontram-se alguns exemplos das correspondências efetuadas entre as categorias e subcategorias Colab e as dimensões da ISO. Para a base Colab não existe a classe Nenhum, pois todas as postagens foram enquadradas em alguma dimensão. As seguintes dimensões ISO não obtiveram nenhuma correspondência nas classes Colab e, portanto, não foram consideradas nos experimentos: Educação, Finanças, Resposta à Incêndios e Emergências, e Telecomunicações. Da mesma forma que na base Twitter, os dados Colab também se apresentam altamente desbalanceados, como pode ser observado na Tabela 3, onde Transportes é a dimensão ISO com maior amostra.

Tabela 4 – Exemplos do mapeamento entre Colab e ISO

Dimensão ISO	**Colab**	
	Categoria	**Subcategoria**
Água e Saneamento	Água e Esgoto	Falta de água
Águas Residuais	Água e Esgoto	Esgoto a céu aberto
Economia	Estabelecimento irregular	Preço abusivo do combustível
Energia	Iluminação e Energia	Falta de energia
Energia	Iluminação e Energia	Lâmpada acesa de dia
Meio Ambiente	Área Rural	Desmatamento Ilegal
Meio Ambiente	Estabelecimento irregular	Emissão de fumaça preta
Meio Ambiente	Meio Ambiente	Poda ou retirada de árvore
Resíduos Sólidos	Aeroporto	Sujeira no aeroporto
Resíduos Sólidos	Limpeza e Conservação	Lixeira quebrada
Saúde	Saúde	Foco de dengue
Saúde	Saúde	Infestação de roedores
Segurança	Estabelecimento irregular	Estabelecimento sem saída de emergência
Segurança	Segurança	Ponto de assalto/roubo
Transportes	Copa do Mundo	Ônibus danificado
Transportes	Pedestres e Ciclistas	Calçada irregular
Transportes	Pedestres e Ciclistas	Faixa de pedestre inexistente

Fonte: Elaborado pela autora.

4.2.2 Pré-processamento

A categorização em dimensões ISO proposta neste artigo parte da premissa que a ordem das palavras dentro da mensagem, estejam elas em maiúsculas ou minúsculas ou que tipo de pontuação é usada, não afeta a tarefa de classificação. Com base nisso, aplicou-se as seguintes técnicas de processamento de texto.

4.2.2.1 Padronização e Filtragem

Todas as mensagens foram padronizadas para caixa baixa. Foram removidos links, pontuação, números (incluindo data e hora) e quaisquer símbolos (incluindo símbolo de *hashtags* e outros caracteres especiais). Além disso, foram retiradas todas as menções à contas. Uma lista de *stopwords* (palavras que não contribuem ou atrapalham a classificação das mensagens) em português foram removidas dos documentos. As *stopwords* desta pesquisa contém principalmente artigos, preposições, conjunções e pronomes.

4.2.2.2 Stemming

Um processo de *stemming* foi executado usando o mesmo *stemmer* aplicado no dicionário. O objetivo do *stemming* é reduzir o número de termos considerados como entrada. Ambos os conjuntos de dados, Twitter e Colab, foram submetidos ao mesmo pré-processamento. A Figura 28 apresenta alguns exemplos de mensagens antes e depois da padronização, filtragem e *stemming*.

4.2.2.3 Tokenização

Os atributos foram extraídos por meio de um processo de tokenização 3-gramas resultando em duas matrizes: $MT_{1950,31688}$, para o Twitter, e $MC_{65066,1089340}$, para o conjunto de dados do Colab. Na Tabela 5 podem ser verificadas algumas métricas dos dois datasets (conjuntos de dados) utilizados, comparando os valores antes e depois do pré-processamento.

Figura 28 – Exemplos de mensagens processadas

Mensagem Original	Mensagem processada
@saudepoa Parabens pela iniciativa. Apoiamos sempre este tipo de projeto. As pessoas lembram e se comovem muito com crianças carentes e doentes,mas esquecem dos idosos, que muitas vezes, são abandonados por suas famílias!	parabens inic apoi sempr tipo projet pesso lembr comov crianc carent doent esquec idos muit vez abandon famil
@EPTC_POA oi, sinaleira não tá funcionando na Getúlio x Ipiranga	oi sinaleir ta funcion getuli ipirang
@Prefeitura_POA Não posso. Não há estacionamento com vagas reservadas pra deficientes no local.	poss estacion vag reserv pra deficient local
@portoalegre24h é o preço que se paga por se jogar tudo que é lixo na rua!	prec paga jog tudo lixo rua
@DMLU_POA Onde posso descartar um forno de microondas?	onde poss descart forn microond
@dmaepoa Quando pagamos o boleto tbm é bem complexo.	pag bolet tbm bem complex
@EPTC_POA. Cratera na curva da Lucas de Oliveira x Neusa G Brizola. Vários automóveis com pneu rasgado	crat curv luc oliveir neus brizol vari automov pneu rasg
@Prefeitura_POA Tem que fazer nos bairro afastados do centro tambem, Rua Itambe, bairro Lami, 3 meses sem iluminação publica	faz bairr afast centr tamb rua itamb bairr lami mes ilumin public

Fonte: Elaborado pela autora.

Tabela 5 – Métricas do Pré-processamento dos Datasets

Dataset	**Palavras**	**Média Palavras por doc**	**Média Caract. por pal.**	**Tokens 3-grams**
Twitter antes	31.201	16,0	4,6	48.420
Twitter depois	18.060	9,3	4,9	31.688
Colab antes	1.677.894	25,8	4,8	1.389.363
Colab depois	992.370	15,3	5,0	1.089.340

Fonte: Elaborado pela autora.

4.2.3 Seleção dos algoritmos de classificação

Com o objetivo de identificar os algoritmos mais adequados dentre os disponíveis na biblioteca Scikit-learn, versão 0.20.2, reduziu-se a dimensionalidade de MT e MC através da utilização do método *TFIDFVectorizer* da Scikit-learn usando a configuração padrão acrescida dos parâmetros:

- $sublinear\_tf = True$: utiliza uma variante do TF-IDF onde TF é substituído por $1 + log(TF)$,
- $ngram\_range = (1, 3)$: a tokenização é feita considerando 1, 2 e 3 gramas,
- $min\_df = 5$: são considerados somente os atributos que ocorrem em pelo menos 5 documentos.

Após o processamento da nova vetorização constatou-se grande redução no número de atributos das novas matrizes: $MT^{a}_{1950,919}$ e $MC^{a}_{65066,44663}$. A redução do tamanho das matrizes facilitará a identificação dos melhores classificadores da Scikit-learn dentre um conjunto maior de algoritmos.

Inicialmente definiu-se a estratégia de seleção de algoritmos para classificação supervisionada considerando: identificar algoritmos de classificação disponíveis na Scikit-learn, definir a estratégia de treino e teste e a métrica de avaliação a ser utilizada. Onze algoritmos foram selecionados considerando sua facilidade de implementação em Python. Inicialmente, aplicou-se os algoritmos somente sobre os dados do Twitter (MT^{a}).

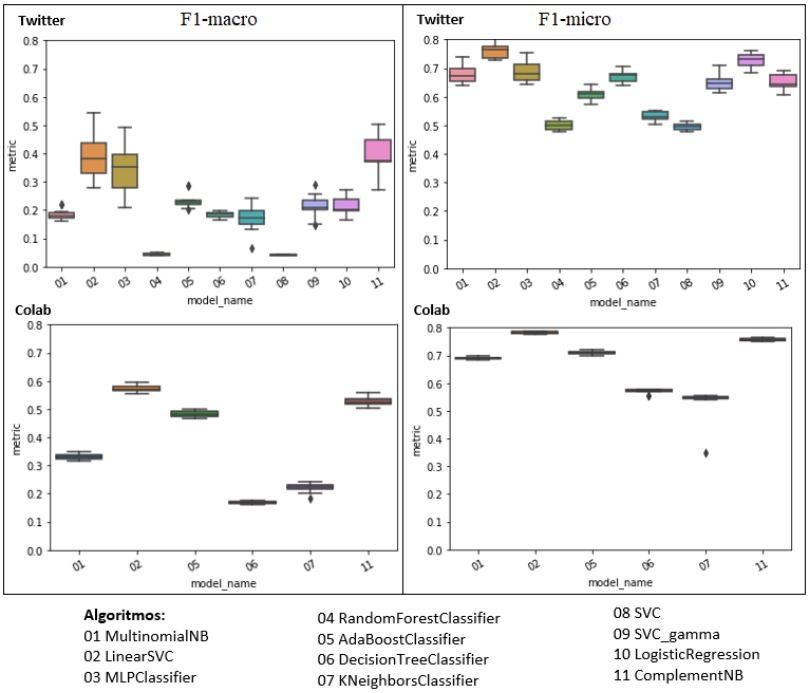
Definiu-se como métricas de avaliação do desempenho as médias de pontuação F1-score macro ($F1^{M}$) e micro ($F1^{\mu}$). A baixa generalidade de algumas classes no conjunto de dados Colab e Twitter justificam o uso de macro médias como o principal direcionador na seleção dos algoritmos a serem utilizados na etapa de seleção de atributos, descrita na Seção 4.2.4.

Os onze classificadores foram testados no Twitter aplicando-se a estratégia CV com 7 *folds*. Utilizou-se a função da Scikit-learn $cross\_val\_score$ que divide os dados repetidamente em conjuntos de treinamento e teste, treina o preditor usando o conjunto de treinamento e calcula as pontuações com base no conjunto de testes para cada *fold* da CV, apresentando como resultado a média de todos os $folds$. Para utilizar esta função é necessário informar os seguintes parâmetros:

- o algoritmo preditor (um dos algoritmos da Tabela 6),
- a matriz termo-documento (X, que corresponde a MT^a e MC^a),
- a métrica desejada (foram utilizadas F1-macro e F1-micro) e
- o número *k-folds* do CV (foram utilizadas 7 *folds*).

Para o conjunto de dados MC^a, manteve-se a mesma estratégia, entretanto devido ao alto tempo de processamento na base Colab foram aplicados apenas os cinco algoritmos com melhor custo benefício no banco de dados do Twitter (tempo de processamento e resultados da F1-macro). Os algoritmos foram aplicados para ambas as bases utilizando as instruções e os parâmetros descritos na Tabela 6.

Figura 29 – Resultados preliminares dos Classificadores



Fonte: Elaborado pela autora.

Analisando os resultados apresentados na Figura 29, é possível verificar que as métricas da base Colab mostram menor variabilidade para os 7 *folds* do que os dados do Twitter. Além disso, as médias

Tabela 6 – Algoritmos Scikit-learn selecionados

#	**Instrução e Parâmetros**
01	**MultinomialNB** (alpha=1.0, class_prior=None, fit_prior=True)
02	**LinearSVC** (C=1.0, class_weight=None, dual=True, fit_intercept=True, intercept_scaling=1, loss='squared_hinge', max_iter=1000, multi_class='ovr', penalty='l2', random_state=None, tol=0.0001, verbose=0)
03	**MLPClassifier** (activation='relu', alpha=0.0001, batch_size='auto', beta_1=0.9, beta_2=0.999, early_stopping=False, epsilon=1e-08, hidden_layer_sizes=(100,), learning_rate='constant', learning_rate_init=0.001, max_iter=200, momentum=0.9, n_iter_no_change=10, nesterovs_momentum=True, power_t=0.5, random_state=None, shuffle=True, solver='adam', tol=0.0001, validation_fraction=0.1, verbose=False, warm_start=False)
04	**RandomForestClassifier** (bootstrap=True, class_weight=None, criterion='gini', max_depth=3, max_features='auto', max_leaf_nodes=None, min_impurity_decrease=0.0, min_impurity_split=None, min_samples_leaf=1, min_samples_split=2, min_weight_fraction_leaf=0.0, n_estimators=200, n_jobs=None, oob_score=False, random_state=0, verbose=0, warm_start=False)
05	**AdaBoostClassifier** (algorithm='SAMME.R', base_estimator=None, learning_rate=1.0, n_estimators=50, random_state=None)
06	**DecisionTreeClassifier** (class_weight=None, criterion='gini', max_depth=5, max_features=None, max_leaf_nodes=None, min_impurity_decrease=0.0, min_impurity_split=None, min_samples_leaf=1, min_samples_split=2, min_weight_fraction_leaf=0.0, presort=False, random_state=None, splitter='best')
07	**KNeighborsClassifier** (algorithm='auto', leaf_size=30, metric='minkowski', metric_params=None, n_jobs=None, n_neighbors=3, p=2, weights='uniform')
08	**SVC** (C=0.025, cache_size=200, class_weight=None, coef0=0.0, decision_function_shape='ovr', degree=3, gamma='auto_deprecated', kernel='linear', max_iter=-1, probability=False, random_state=None, shrinking=True, tol=0.001, verbose=False)
09	**SVC** (C=1, cache_size=200, class_weight=None, coef0=0.0, decision_function_shape='ovr', degree=3, gamma=2, kernel='rbf', max_iter=-1, probability=False, random_state=None, shrinking=True, tol=0.001, verbose=False)
10	**LogisticRegression** (C=1.0, class_weight=None, dual=False, fit_intercept=True, intercept_scaling=1, max_iter=100, multi_class='warn', n_jobs=None, penalty='l2', random_state=0, solver='warn', tol=0.0001, verbose=0, warm_start=False)
11	**ComplementNB** (alpha=1.0, class_prior=None, fit_prior=True, norm=False)

Fonte: Elaborado pela autora.

F1-macro da base da Colab são ligeiramente superiores. Os algoritmos com maior pontuação são *Complement Naïve Bayes* (CNB) (RENNIE et al., 2003), alcançando F1-macro de 0,40 e 0,53 para dados do Twitter e Colab, respectivamente, e *Linear Support Vector Classification* (LSVC) (FAN et al., 2008), obtendo 0,39 e 0,57.

4.2.3.1 Complement Naïve Bayes

O classificador CNB foi projetado para corrigir as "suposições severas" feitas pelo classificador *Naïve Bayes Multinomial* original empregando soluções heurísticas simples que mantém o algoritmo rápido e fácil de se implementar (RENNIE et al., 2003). De acordo com SCIKIT-LEARN (2019), CNB é particularmente adequado para conjuntos de dados desequilibrados.

Com respeito a abordagem para o cálculo dos pesos do modelo, ao invés de calcular a probabilidade de ocorrência de uma palavra em uma classe, como na versão tradicional do Naïve Bayes, calcula-se a probabilidade desta palavra ocorrer em outras classes. Desta forma, o CNB usa estatísticas do complemento de cada classe para calcular os pesos do modelo (SCIKIT-LEARN, 2019).

4.2.3.2 Linear Support Vector Classification

O classificador LSVC faz parte da família de classificadores SVM da biblioteca Scikit-learn e utiliza $kernel = linear$ (PEDREGOSA et al., 2011). Está implementado nos termos da biblioteca LIBLINEAR descrita por Fan et al. (2008). Segundo SCIKIT-LEARN (2019) LSVC possibilita maior flexibilidade na escolha de penalidades e funções de perda.

4.2.4 Seleção de Atributos com Wrapper

Buscando a melhora dos resultados dos dois algoritmos nomeados na Seção 4.2.3 foram aplicadas técnicas para seleção de atributos sobre as matrizes termo-documento originais (MT e MC), utilizando o conceito de *wrapper*: técnicas de seleção de atributos foram testadas uma a uma em combinação com os dois melhores preditores (CNB e LSVC). Em todos os experimentos de seleção de atributos manteve-se os demais parâmetros dos dois algoritmos com a configuração padrão

da Scikit-learn. Os dados foram treinados e testados usando CV com *k-fold*=7. As métricas F1-macro foram calculadas ao longo de uma série de iterações buscando pela melhor configuração (técnica de seleção de atributos + número de atributos + preditor) dentro das combinações testadas. Um procedimento de busca $Sr(Tr, Fs)$ foi desenvolvido, onde Tr representa um limite do número de atributos ou o número de documentos que contém o atributo, e Fs é a função responsável por calcular o melhor Ts (que resulte em um maior F1-macro).

Dois grupos de métodos foram investigados para extrair e selecionar características relevantes. O primeiro engloba métodos que não consideram os rótulos de classe: *Document Frequency* (DF, Frequência em Documentos) e *Maximum Number of Features* (MNF, Número Máximo de Atributos). O segundo grupo consiste em métodos estatísticos que levam em consideração os rótulos de classe: Chi^2, ANOVA F-test (*Analysis of Variance*) e *Mutual Information* (MI). Além das técnicas de seleção de atributos, os experimentos consideraram duas formas diferentes de calcular os pesos da representação Bag of Words (BOW) que denominaremos "vetorizadores": (a) contagens da ocorrência dos termos nos documentos (*Countvectorizer*) e (b) pesos TF-IDF (*TFIDFvectorizer*). Os dois métodos de vetorização dos documentos foram usados com codificação (*encoding*) "latin1" e a tokenização considerou 1, 2 e 3 gramas ($ngram\_range = (1, 3)$). Para *TFIDFvectorizer* a normalização utilizada foi a padrão $l2$ ($norm = l2$), que corresponde a distância euclidiana (MANNING; RAGHAVAN; SCHüTZE, 2009). Todos os métodos de seleção de atributos relatados a seguir foram executados utilizando CV com *k-folds*=7 e utilizando a função $cross\_val\_score$ da Scikit-learn. Cada experimento é detalhado nas subseções a seguir.

4.2.4.1 Frequência de Documentos (DF)

Os atributos são reduzidos mantendo-se apenas os termos mais frequentes (acima de um determinado limite em função do número ou porcentagem de documentos contendo o atributo). Esse é um método de seleção de atributos não supervisionado, pois não considera rótulos de classe. Essa técnica assume que termos raros são menos informativos para algoritmos de aprendizado e que palavras frequentes contribuem mais(AGARWAL; MITTAL, 2014).

Primeiramente, Tr foi usado como o número de documentos e executou-se Sr iterando $Tr \subseteq [0, 19]$. Na Scikit-learn utiliza-se o parâmetro min_df recebendo os valores de Tr ao executar os vetorizadores.

Para o conjunto de dados do Twitter, obtemos F1-macro de 0,47 aplicando LSVC em uma representação BOW com contagem da ocorrência dos tokens (*Countvectorizer*) e considerando $Tr = 3$. Para Colab, a melhor configuração obtida foi 0,58 também com LSVC e $Tr = 2$, mas rodando sobre uma BOW com pesos TF-IDF (*TFIDFvectorizer*).

O comportamento de DF ao longo da série de experimentos é apresentado na Figura 30. Os dados do Twitter apresentam uma diminuição mais pronunciada no desempenho quando aumenta Tr, o que está relacionado ao menor volume de dados em comparação ao Colab. Como o conjunto de dados do Colab tem um volume muito maior, também foi executado Sr com o parâmetro Tr contendo a porcentagem de documentos e variando de 0,01% a 2%, entretanto, os resultados não superaram o primeiro experimento. Para aplicar essa técnica, é importante considerar a distribuição das categorias na amostra, ou seja, Tr deve ser menor que a categoria com menor número ou percentual de amostras dentro do conjunto de dados.

4.2.4.2 Numero Máximo de Features (MNF)

Nesse caso, Tr corresponde ao número de atributos (*features*) a serem selecionados de uma lista ordenada da maior frequência dos termos para a menor. Para aplicar esse conceito nos vetorizadores da Scikit-learn utiliza-se o parâmetro $max\_features$ recebendo os valores de Tr.

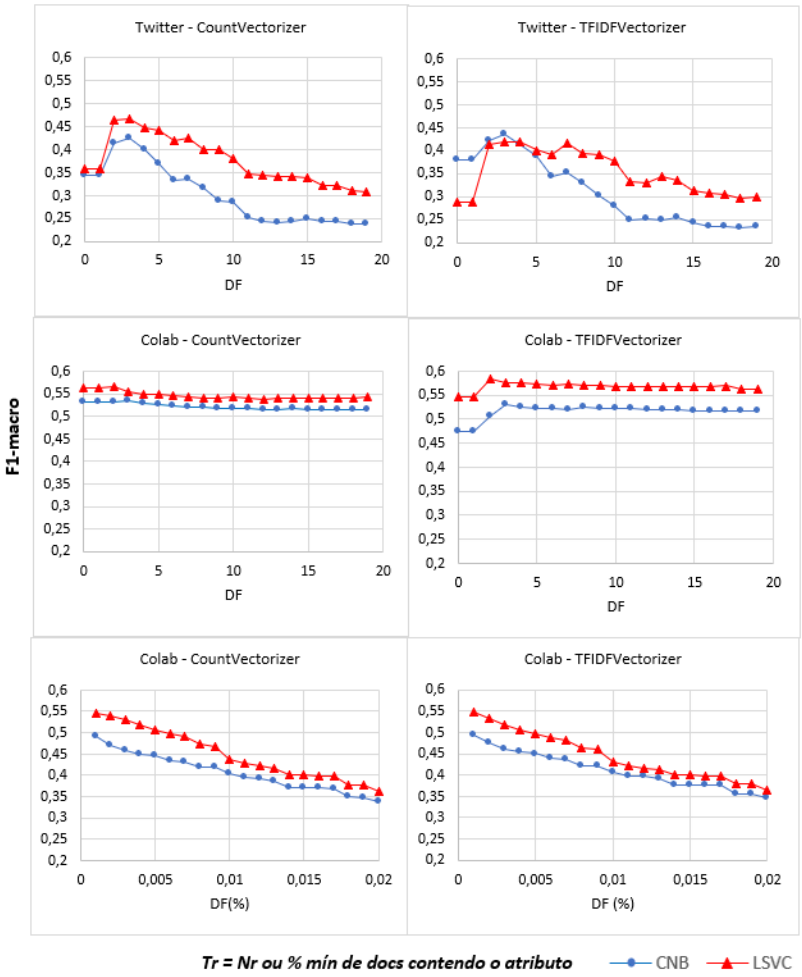
A melhor configuração para o conjunto de dados do Twitter alcançou 0,48 com LSVC, *Countvectorizer* e Tr =8.188. O LSVC também obteve os resultados mais altos da MNF (0,58) no conjunto de dados Colab, com Tr =237.107. A variação de Tr e os respectivos resultados da F1-macro são apresentados na Figura 31.

4.2.4.3 Métodos Estatísticos

Nas abordagens com métodos estatísticos, Tr corresponde ao número de atributos a ser selecionado por cada função Fs. Neste trabalho foram aplicadas às seguintes funções estatísticas disponíveis no modulo $feature\_selection$ da Scikit-Learn para classificação:

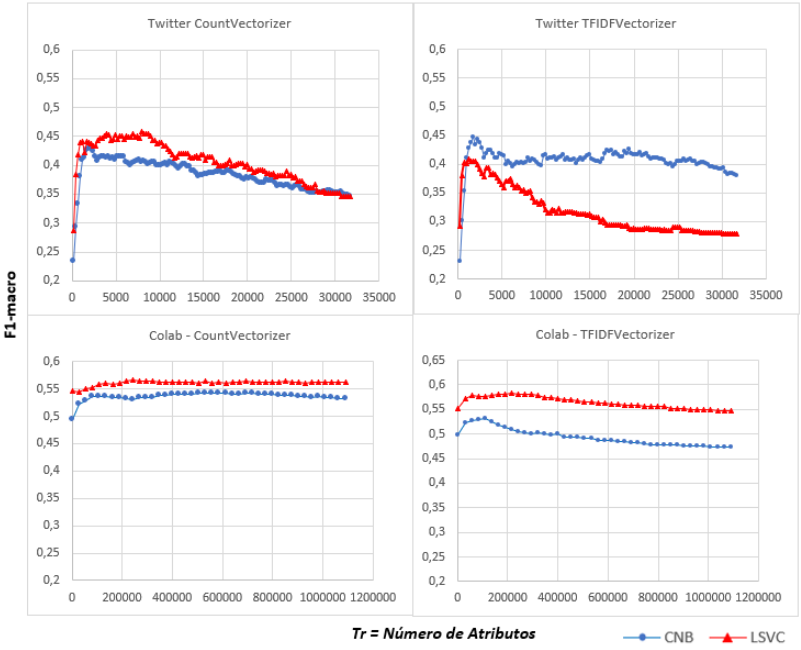
- Chi^2 ou X^2 (LIU; SETIONO, 1995; AGARWAL; MITTAL, 2014) usando o método chi^2;
- ANOVA (*Analysis of variance*) F-Value (KUMAR et al., 2015),

Figura 30 – Procedimento *Sr(Tr,DF)* e resultados F1-macro



Fonte: Elaborado pela autora.

Figura 31 – Procedimento *Sr(Tr, MNF)* e os resultados F1-macro



Fonte: Elaborado pela autora.

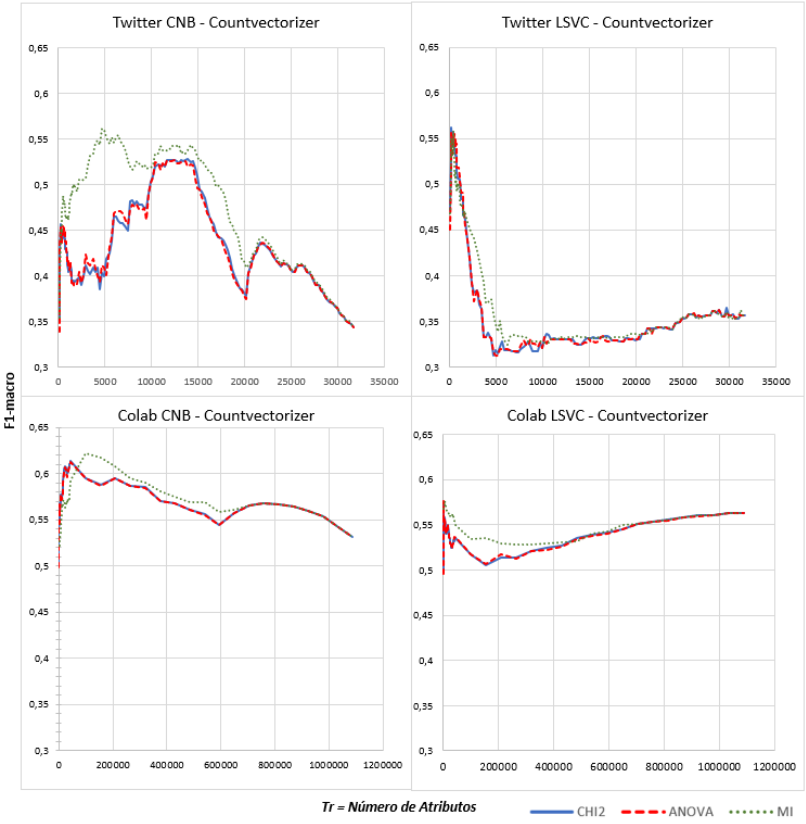
usando o método *f-classif*,

- MI (*Mutual Information*) (KRASKOV; STÖGBAUER; GRASSBERGER, 2004; HOQUE; BHATTACHARYYA; KALITA, 2014) utilizando o método *mutual_info_classif*.

A implementação da seleção de atributos usando os três métodos estatísticos mencionados foi efetuado utilizando-se a função $SelectKBest$ da Scikit-learn que implementa um "seletor". Esta função recebe como parâmetro a função estatística a ser utilizada ($score\_func$) e o número dos melhores atributos a serem selecionadas (k), onde k recebe os valores de Tr. Após a configuração do seletor ele é aplicado sobre MT e MC e seus respectivos rótulos, resultando em novas matrizes MT^b e MC^b que possuem somente os melhores atributos selecionados através da respectiva função estatística. Estas novas matrizes são então submetidas a função $cross\_val\_score$ juntamente com os demais parâmetros da respectiva iteração.

O desempenho geral dos classificadores, aplicando as três técnicas de seleção por meio de iterações do número de atributos (Tr) a ser selecionado para cada configuração, é apresentado na Figura 32, onde estão demonstrados os resultados obtidos nos três procedimentos usando o vetorizador *CountVectorizer*: $Sr(Tr, X^2)$, $Sr(Tr, ANOVA)$, $Sr(Tr, MI)$. Observa-se que MI apresenta desempenho superior ao longo da maioria dos intervalos de Tr. Os comportamentos de $ANOVA$ e Chi^2 revelaram-se muito semelhantes.

Figura 32 – Sr com métodos estatísticos e resultados F1-macro



Fonte: Elaborado pela autora.

5 RESULTADOS E DISCUSSÃO

5.1 RESULTADOS

A Tabela 7 mostra os desempenhos dos modelos usando CNB e LSVC e considerando as melhores configurações obtidas no estágio anterior. Calculou-se $F1^M$ e $F1^\mu$ para cada uma das melhores configurações aplicando CV com 7 *folds* por meio de *stratified cross validation* (validação cruzada estratificada), buscando preservar a proporção das categorias. Neste cálculo empregou-se a função $StratifiedKFlods$ da Scikit-learn como um parâmetro adicional à função $cross\_val\_score$.

Tabela 7 – Melhores configurações para classificação ISO

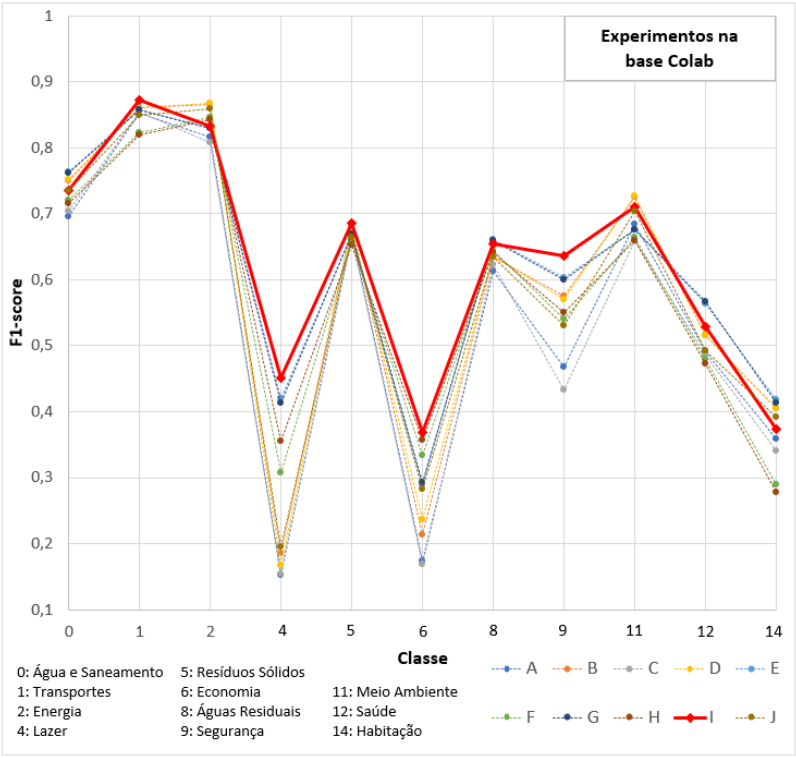
Base	Exp	Seleção de atributos	Clf	Vetorizador	Atributos usados	Redução atributos	$F1^M$	$F1\mu$
Colab	A	MFN	CNB	Count	612.340	44%	0,5429	0,7551
	B	MFN	LSVC	TFIDF	214.840	80%	0,5830	0,7853
	C	min DF = 3	CNB	Count	95.658	91%	0,5358	0,7550
	D	min DF = 2	LSVC	TFIDF	237.107	78%	0,5837	0,7860
	E	Chi2	CNB	Count	44.340	96%	0,6136	0,7751
	F	Chi2	LSVC	Count	850	99,9%	0,5734	0,7566
	G	ANOVA	CNB	Count	44.340	96%	0,6127	0,7749
	H	ANOVA	LSVC	Count	700	99,9%	0,5769	0,7533
	I	**MI**	**CNB**	**Count**	**99.340**	**91%**	**0,6225**	**0,7852**
	J	MI	LSVC	Count	6.340	99%	0,5761	0,7719
Twitter	L	MNF	CNB	TFIDF	1.688	95%	0,4419	0,6656
	K	MNF	LSVC	Count	8,188	74%	0,4820	0,7810
	M	min DF = 3	CNB	TFIDF	1.711	95%	0,4357	0,6641
	N	min DF = 3	LSVC	Count	1.711	95%	0,4674	0,7590
	O	Chi2	CNB	Count	13.938	56%	0,5277	0,7359
	P	Chi2	LSVC	Count	220	99%	0,5617	0,7154
	Q	ANOVA	CNB	Count	11.688	63%	0,5271	0,7451
	R	ANOVA	LSVC	Count	188	99%	0,5562	0,6651
	S	**MI**	**CNB**	**Count**	**4.688**	**85%**	**0,5615**	**0,7051**
	T	MI	LSVC	Count	470	99%	0,5587	0,7928

Exp: Experimentos; **Clf**: Algoritmo de Classificação;
Fonte: Elaborado pela autora.

Considerando a métrica F1-macro, o melhor cenário em Colab é I com as maiores médias macro e micro. Quando o resultado é ob-

servado individualmente nas classes, a Figura 33 também mostra que I corresponde ao melhor desempenho. Próximos aos resultados atingidos por I, temos E e G, também com CNB usando um menor número de atributos selecionados através de Chi^2 e ANOVA, respectivamente. O custo computacional maior da MI é tolerável para o tamanho do conjunto de dados, portanto, manteve-se a opção de CNB + MI que corresponde ao cenário I.

Figura 33 – Resultados Colab abertos por dimensão ISO

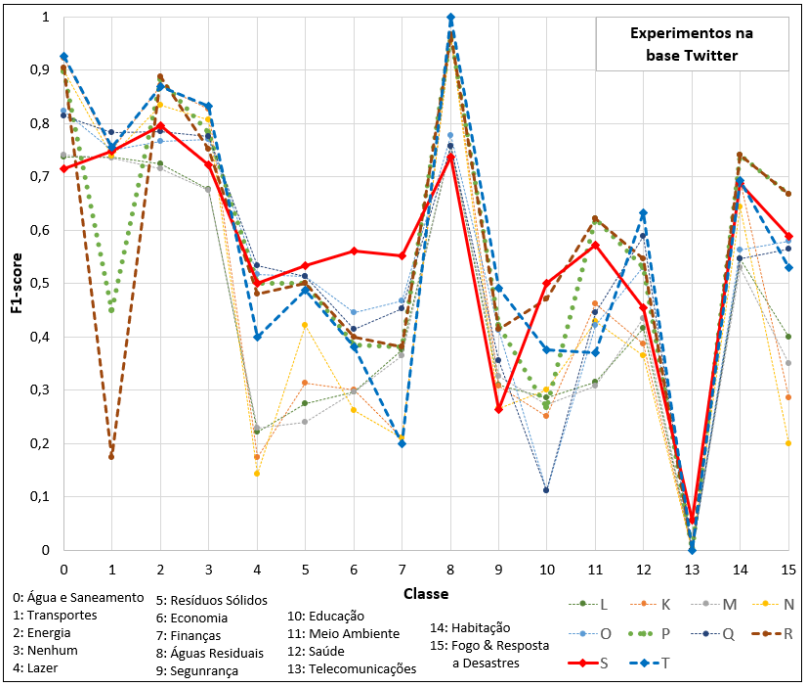


Fonte: Elaborado pela autora.

No Twitter, os experimentos P, R, S e T alcançaram as maiores performances, em torno de 0,56 para $F1^M$. O desempenho individual das dimensões ISO na Figura 34 foi avaliado para determinar a melhor configuração, onde podemos ver resultados que variam muito por classe. Rejeitou-se os cenários P e R devido aos baixos valores para Transportes, que é uma das dimensões com maior volume no corpus. A decisão

sobre S ou T foi baseada principalmente no número de categorias com maior F1-score, excluindo da avaliação a categoria Telecomunicações por ter volume muito baixo, gerando um desempenho muito ruim da classe. Como resultado, determinou-se S como a melhor opção.

Figura 34 – Resultados Twitter abertos por dimensão ISO

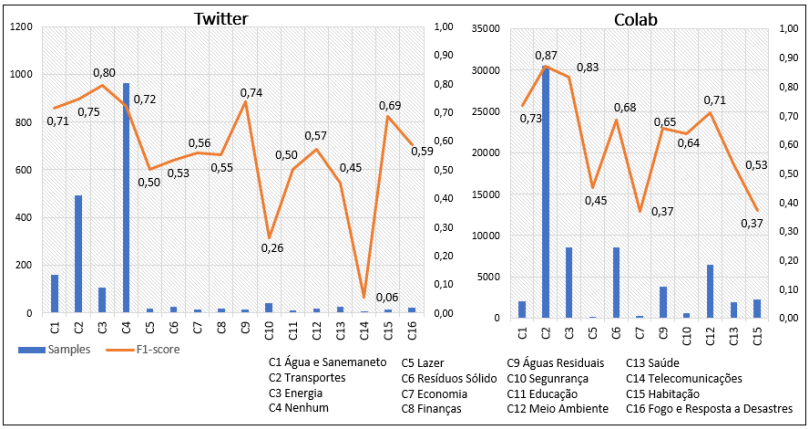


Fonte: Elaborado pela autora.

O desempenho na base Colab é melhor que na base Twitter na maioria das dimensões comuns, como pode ser visto na Figura 35. Isso se deve ao volume maior da base Colab e também a um menor número de classes ISO: são consideradas 11 classes, enquanto o Twitter tem 16 e 3% do tamanho da base Colab.

No Apêndice A, é possível verificar mais detalhes sobre os atributos selecionados nos MCS vencedores, relativos aos experimentos I e S.

Figura 35 – Desempenho dos MCS selecionados nas dimensões ISO



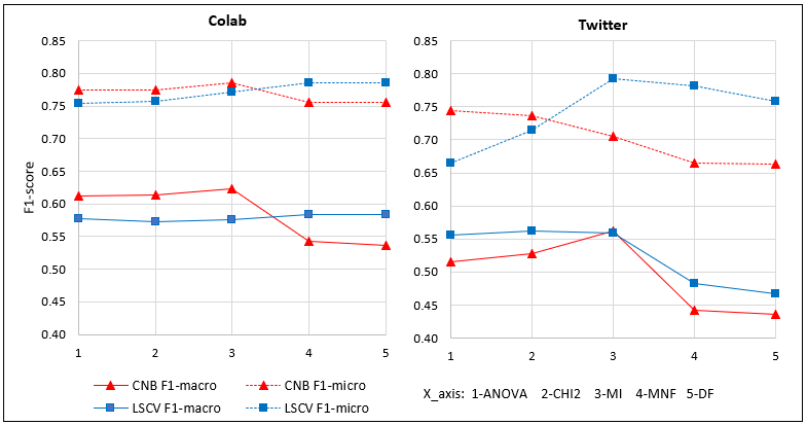
Fonte: Elaborado pela autora.

5.2 DISCUSSÃO

A métrica F1-macro foi usada na comparação dos desempenhos dos modelos em conjunto com o F1-micro e os resultados internos das classes. A figura 36 mostra os resultados gerais focando nas cinco técnicas de seleção de atributos utilizadas em cada um dos dois algoritmos, conforme detalhado na Tabela 7.

Twitter: DF e MNF apresentaram valores baixos para macro médias, mas para LSVC apresentam um resultado micro melhor do que ANOVA ou CHI^2, mostrando que não devem ser técnicas descartadas à primeira vista, uma vez que sua execução é mais rápida que os testes estatísticos. Isto é reforçado pelos resultados na base Colab com LSVC, onde DF e MNF tem desempenho muito próximos dos métodos estatísticos. As macro médias do Twitter para LSVC e CNB têm um comportamento próximo ao longo dos cinco métodos, entretanto para micro médias elas apresentam grandes diferenças, com LSVC reagindo melhor a MI, DF e MNF e CNB melhor a ANOVA e Chi^2. MI traz os resultados mais altos em macro médias para ambos os algoritmos e, no caso do LSVC, resulta em um F1-micro mais alta também. ANOVA e Chi^2 com LSVC também representam boas opções considerando F1-macro. Similarmente à base Colab, no Twitter os resultados com CNB + MI foram ligeiramente melhores do que as demais configurações, entretanto menos claros do que no Colab. A partir da avaliação do

Figura 36 – Seleção de Atributos vs. Algoritmos



Fonte: Elaborado pela autora.

desempenho individual das classes no Twitter foi possível determinar a opção CNB + MI como o MCS que este trabalho objetivava.

Colab: O comportamento dos dois algoritmos sobre os diferentes métodos de seleção de atributos é mais próximo do que no Twitter. Para LSVC, técnicas como MNF ou DF alcançam resultados similares aos métodos estatísticos, considerando macro médias, além de apresentarem cerca de 0,02 pontos acima de Chi^2 e ANOVA para o F1-micro. Por outro lado, o CNB funciona melhor com as funções estatísticas: MI obteve uma pequena superioridade em comparação a Chi^2 e ANOVA, porém teve um custo computacional bem maior, o que é uma questão a ser investigada em trabalhos futuros. A conclusão de que CNB + MI compõem a melhor configuração para o Colab coincide com o experimento que possui a melhor F1-macro: *I*. Quando olhamos para o desempenho interno de *I* contra os outros experimentos, é bastante claro que é a melhor opção, como demonstrado na Figura 33.

Avaliando-se os resultados do MCS definido para as duas bases e apresentado na Figura 35, é possível compreender os resultados obtidos para cada dimensão da cidade. As classes com melhor desempenho na base Colab foram Transportes, Água e Saneamento, Energia e Meio Ambiente. A classe Água e Saneamento chama a atenção, pois têm um pequeno volume de dados rotulados, cerca de 3,1% do corpus, mas atinge 0,73, o que demonstra que a classe possui atributos fortemente conectados a ela, por exemplo "água". O Twitter também tem

as classes Energia, Transportes, Água e Saneamento com as mais altas pontuações acrescido da classe Águas Residuais, a qual figura na terceira posição. Transportes igualmente corresponde a dimensão ISO com maior amostra no Twitter. Como ocorreu para a base Colab, algumas classes com pouquíssimas amostras obtiveram uma boa pontuação, como Energia e Águas Residuais, fato que está relacionado aos atributos muito representativos das classes contendo, por exemplo, os *tokens* de "luz" e "esgoto", respectivamente. A classe Telecomunicações deverá ser revisada, pois as amostras são muito poucas, produzindo resultados ruins para esta dimensão. As perguntas propostas para esta pesquisa (PP) são discutidas a seguir.

PP1 - Como classificar as interações de RSO nas dimensões da ISO 37120?

A metodologia utilizada nesta dissertação para coletar, processar e classificar as manifestações nas RSO relacionadas às dimensões da ISO 37120, considera abordagens destacadas por vários autores referenciados na fundamentação teórica. A partir deste embasamento, foi proposto e executado o fluxo da Figura 21 composto por três grandes grupos de processos que respondem a PP1: a confecção de uma base rotulada, o pré-processamento das mensagens e a classificação das mesmas usando o método *wrapper* de seleção de atributos.

Um aspecto importante na elaboração da base rotulada é como selecionar mensagens que se enquadrem nas classes. O desenvolvimento do dicionário de termos da ISO facilitou essa tarefa. A partir deste dicionário foi possível localizar mensagens do Twitter com maior probabilidade de se enquadrar em uma das classes, bem como mensagens sem nenhuma ocorrência do dicionário possuem maior probabilidade de pertencer a classe "Nenhum". Uma vez pré-selecionadas, estas mensagens foram classificadas manualmente compondo o que é denominado "Base rotulada", fundamental para abordagens de aprendizado supervisionado.

Com respeito às técnicas de pré-processamento utilizadas, cabe ressaltar que o objetivo de aplicar as mesmas é facilitar a classificação através da padronização de mensagens, remoção de *stopwords* e *stemming*, permitindo reduzir o número de tokens considerados. Por outro lado, a tokenização 3-gramas possibilita a identificação de atributos com até três palavras que muitas vezes são mais representativos da classe do que se considerarmos as palavras da expressão individualmente, por exemplo "área verde", "população em situação de rua", "fios

de alta tensão".

No que tange à tarefa de classificação, a utilização da biblioteca Scikit-learn e da linguagem Python permitiu o teste de vários algoritmos de classificação supervisionada, propiciando a prática das diferentes famílias de algoritmos discutidas na seção 2.5 dentro do contexto de dados existentes. O foco do trabalho foi executar estes algoritmos através do método *wrapper* para seleção de atributos que melhorassem o resultado da F1-macro. Desta forma, não foram exploradas variações nos parâmetros dos algoritmos, o que é escopo para trabalhos futuros. Com respeito a escolha do melhor MCS, utilizou-se a F1-macro como principal métrica direcionadora, e para decidir entre valores muito próximos considerou-se também a F1-micro e o F1-score de cada classe. Não foi avaliada a significância estatística nas comparações de cada experimento, o que deve ser endereçado em trabalhos futuros.

Apesar do processo e técnicas utilizadas para definir o melhor MCS responderem a PP1, os modelos apresentaram limitações para classificar categorias com poucas amostras nos conjuntos de dados. Este é o caso, por exemplo, da categoria Telecomunicações (nos dados do Twitter) onde os modelos não conseguiram classificar as mensagens. Como pode ser visto na Figura 35, há uma alta variação entre os melhores e os piores resultados das classes (cerca de 0,5 de variação para ambos conjuntos de dados). Além disso, 36% das categorias do Colab alcançaram F1-score abaixo de 0,6 e para o Twitter isso corresponde a 60% das classes. Esta limitação pode ser endereçada no nível dos dados e dos algoritmos. No nível dos dados pode-se aplicar métodos para lidar com classes desbalanceadas, usando técnicas de reamostragem e/ou aprofundar os métodos de seleção de atributos. Já no nível de algoritmos, pode-se focar em *cost sensitive learning* (aprendizado sensível ao custo), que procura minimizar os erros de classificação, ou métodos *embedded*, que combinam diferentes algoritmos de aprendizado.

PP2 - Quais são os desafios técnicos para estender essa classificação a outros modelos de indicadores de cidades?

Com relação à segunda pergunta da pesquisa PP2, os principais desafios técnicos para incorporar outros modelos de Cidades Inteligentes estão relacionados principalmente à correspondência das novas dimensões com as existentes. Os classificadores da ECMC são dependentes do modelo de Cidades Inteligentes, isso significa que para adicionar novos, como por exemplo os modelos das Nações Unidas ou o Europeu, será necessário executar novamente o processo descrito na Figura 21.

Além disso, mudanças em modelos de Cidades Inteligentes já contemplados na ECMC, como é o caso de alteração do escopo de uma classe ou adição/remoção de categorias, impactam o serviço de classificação de forma similar à incorporação de um modelo totalmente novo.

Normalmente, os modelos de Cidades Inteligentes têm uma alta sobreposição de dimensões/categorias. Por exemplo, a categoria Transportes existe em quase todos, mas em alguns esta dimensão se encontra dentro de Infraestrutura ou de Mobilidade. Consequentemente, uma boa parte das tarefas de adaptação se refere a mapear os termos do dicionário e os documentos rotulados para o conjunto de classes do novo modelo. O benefício de ter ISO como o primeiro modelo é o fato do mesmo ser um dos modelos conhecidos com maior número de dimensões, portanto, o processo de correspondência de um modelo com maior número de classes para um menor tende a ser mais simples.

Para responder à PP2, propõe-se os passos descritos no Algoritmo 1. Ele representa o fluxo para estender o serviço de classificação incorporando um novo modelo ($Model2$) a partir de um modelo existente ($Model1$). Para cada nova classe checa-se a correspondência ($Match$) para as classes do modelo existente ($classes1$). Obtém-se o dicionário ($ExtractDic$) das classes existentes comparando-o ao escopo da nova classe, verificando novas palavras se necessário ($GetClassWords$), se existem novas palavras as mesmas são adicionadas ao dicionário ($AddDict$). Nesse caso, novos documentos são selecionados no corpus ($GetDocsWithWords$) e classificados ($LabelDocs$). O *wrapper* é executado ($BestSLM$) adicionando os novos documentos rotulados (D) à base geral rotulada e o melhor MCS é determinado para o novo modelo. O MCS definido é adicionado ao serviço de classificação ($AddClassifier$). No caso de uma dimensão totalmente nova, novas palavras são definidas sem usar qualquer dicionário existente como ponto de partida, novos documentos são selecionados a partir das palavras definidas e, então, são rotulados ($LabelDocs$). Se o escopo da nova classe for o mesmo das classes correspondentes, é possível usar os documentos rotulados existentes ($GetLabeledDocs$).

Algoritmo 1: EmbodyModel(Model2)

```
Result: MCS para o novo modelo de Cidade Inteligente a
        ser adicionado ao serviço de Classificação
Model1 ← ISO;
D ← [ ];
forall class in Model2 do
    classes1 ← Match(class, Model1);
    if classes1 then
        dic1 ← ExtractDic(classes1);
        dic2 ← GetClassWords(class, optional dic1);
        if dic2 > dic1 then
            docs ← GetDocsWithWords(dic2 - dic1);
            D ← D + LabelDocs(docs);
            AddDict(dic2 - dic1, class);
        else
            D ← D + GelLabeledDocs(classes1)
        end
    else
        dic2 ← GetClassWords (class);
        docs ← GetDocsWithWords(dic2);
        D ← D + LabelDocs(docs);
        AddDict(dic2, class);
    end
    AddClassifier(BestSLM(D, Model2))
end
```

É importante entender algumas particularidades da cidade que utiliza o serviço de classificação que podem influenciar os resultados. As categorias mais discutidas podem variar dependendo da localidade, das características geográficas, das condições de desenvolvimento humano e de alterações na qualidade do serviço atual. Além disso, o volume de mensagens atribuídas a uma categoria pode ser motivado pela atividade dos perfis oficiais, como é o caso de contas específicas da cidade de Porto Alegre como @eptc_poa e @dmaepoa, que são bastante ativas.

PP3 - Qual seria uma proposta de serviço de classificação viável e suas aplicações nas demandas da cidade?

A ECMC apresentada na Figura 20 canaliza a voz da população que usa RSO para comunicar suas opiniões e reclamações, categorizando-

as em dimensões dos vários modelos de cidades. Desta forma, visa facilitar a comunicação entre cidadãos e prestadores de serviços e contribuir na transformação das cidades em Cidades Inteligentes.

Há vários sistemas de serviços municipais que podem integrar as informações classificadas, tais como: sistemas de decisão de suporte governamental, sistemas de reclamações do cidadão, painéis de comunidades, departamentos de polícia, empresas de transporte, produtores culturais, agências ambientais e empresas de recicladores. Identificou-se quatro grupos de usuários principais do serviço de classificação proposto aqui, como segue:

- Gestores de Serviços Urbanos: representam não só entidades governamentais, como prefeitos e vereadores, mas quaisquer atores envolvidos nos serviços públicos e privados que de alguma forma sejam capazes de impactar a questão encaminhada pelos cidadãos em suas mensagens.

- Cidadãos e Comunidades: os cidadãos em geral estão interessados no que acontece próximo a eles e o fato de ter acesso às manifestações de outros cidadãos dentro da mesma dimensão que lhes interessa no momento, pode colaborar na promoção de ações cívicas e no fortalecimento da conscientização e engajamento dos cidadãos. As comunidades representam Organizações Não Governamentais (ONGs) e qualquer tipo de organização de pessoas por uma causa, como por exemplo, Associação de Pessoas com Deficiência.

- Negócios e empresas: atores envolvidos na dinâmica dos negócios locais têm interesse na voz da população para identificar oportunidades de mercado ou entender riscos relacionados a seus negócios. Alguns exemplos: reclamações sobre a falta de eventos culturais específicos podem motivar produtores de eventos; investidores de futuros empreendimentos podem se interessar por menções em diversas dimensões como segurança, economia, meio ambiente, etc.

- ISO e outras empresas que desenvolvem *ranking* de cidades: a ISO possui um processo de certificação para o padrão 37120. As cidades solicitantes devem enviar anualmente informações oficiais para manter a certificação. A voz da população nas redes sociais conectada à cada dimensão pode alertar para possíveis incoerências entre números oficiais e a manifestação cidadã. Além dessa aplicação, há empresas promovendo *ranking* de cidades,

como *World's Best Cities*[1] e o ranking da Mercer[2]. Empresas como estas geralmente vendem consultoria sobre melhores práticas para atrair investimentos ou para melhor explorar o potencial natural da cidade e torná-la mais atraente. Essas consultorias poderiam usar essa informação como um indicador digital do pulso da cidade nos tópicos específicos.

A viabilidade de um serviço de classificação está fortemente relacionada ao processo de determinação do MCS discutido neste trabalho, uma vez que representa o componente central e mais complexo do *framework*. O volume de informações não deve representar um problema já que para Porto Alegre, cidade com 1,4 milhão de pessoas, o volume de mensagens coletadas é facilmente gerenciável: coletamos 50 mil tweets em 6 meses, seguindo 15 contas. Nem todos os tweets seriam classificados apenas aqueles que mencionam as contas seguidas, cerca de 7.000, o que corresponde a 39 mensagens por dia em média. Todos os pontos explicitados nesta seção respondem a PP3 e demonstram a viabilidade do serviço de classificação.

[1] https://www.bestcities.org/rankings/worlds-best-cities/
[2] https://mobilityexchange.mercer.com/city-benchmarking-and-consulting

6 CONCLUSÕES

Este estudo abordou a classificação de tópicos em conjuntos de dados coletados das redes sociais Twitter e Colab. A principal questão desta pesquisa era a determinação de um MCS viável para classificar as mensagens dos cidadãos em dimensões de Cidades Inteligentes usando a ISO 37120 como primeiro modelo. Para tanto, utilizou-se uma abordagem *wrapper* considerando cinco métodos de seleção de atributos que foram testados através de dois algoritmos de aprendizado de máquina (CNB e LSVC) que tiveram melhor desempenho dentre os onze inicialmente testados. Como os conjuntos de dados eram extremamente desbalanceados, a avaliação do classificador foi feita usando principalmente as médias globais de F1-score (F1-macro). F1-micro e o F1-score de cada classe também foram avaliados. Para ambos os conjuntos de dados, o melhor MCS foi CNB treinado com atributos selecionados através da técnica MI. Os classificadores alcançaram F1-macro de 0,62 para Colab e 0,56 para Twitter e F1-micro de 0,79 para Colab e 0,71 para Twitter. As dimensões ISO que obtiveram os melhores desempenhos pelos classificadores em ambos os conjuntos de dados foram: Água e Saneamento, Transportes, Energia. Especificamente para o Twitter, também podemos destacar as categorias Águas Residuais, Habitação, e para o Colab Meio Ambiente e Resíduos Sólidos.

No Twitter, o melhor MCS não se apresentou de forma tão clara quanto no Colab, fato relacionado ao tamanho da base do Twitter que era muito menor, além de conter cinco classes a mais que a base Colab. Mesmo com essa falta de clareza nos resultados do Twitter, CNB com MI alcançou uma das melhores macro médias, e é o melhor modelo considerando F1-score individual das classes. Como MI exige um alto custo computacional em comparação aos demais métodos, a opção por ele deve sempre considerar o tamanho do conjunto de treinamento, por exemplo no caso de desenvolver um serviço *wrapper* para automatizar o processo de seleção de atributos. Os processos de seleção utilizados aqui buscam a otimização dos classificadores selecionados, que é a essência da abordagem *wrapper*.

Este trabalho envolveu questões amplas relacionadas aos dados das RSO e às diversas dimensões das cidades. Desta forma, é importante esclarecer pontos pertinentes que não foram investigados:

a) Com respeito ao processamento do texto das RSO e da recuperação e classificação da informação, não foram aplicadas técnicas de reamostragem ou técnicas avançadas de PLN como POS (*part-*

of-speech, partes do discurso). Assim, não se considera a interdependência dos termos nas mensagens. Além disso, a classificação aplicada considera que cada mensagem se enquadra somente a uma dimensão, não avaliando um grupo menor de mensagens que se referem a mais de uma categoria (classificação *multilabel*).

b) No âmbito do *framework* proposto, questões de arquitetura da estrutura do serviço de classificação relacionadas ao armazenamento e à publicação dos resultados na Web não foram escopo desta dissertação. O serviço de classificação utilizando o modelo desenvolvido não foi implementado dentro de uma estrutura de API (*Application Programming Interface*). Técnicas de visualização dos resultados para facilitar a compreensão dos dados pelos diferentes atores da cidade também não foram exploradas.

c) Questões relacionadas à segurança e privacidade não foram investigadas profundamente. Acredita-se que a privacidade não seria um problema visto que, apesar de serem contas públicas, remove-se da mensagem a menção a outros perfis. Entretanto, há mudanças em curso nas políticas de privacidade que podem exigir uma anonimização mais profunda das mensagens. Além disso, em tempos de proliferação de *fake news* será válido em etapas futuras avaliar meios de minimizar a classificação de mensagens deste tipo, encontrando critérios para determinar a reputação dos perfis que se manifestam, por exemplo.

A metodologia desenvolvida na classificação das mensagens de RSO em dimensões ISO pode ser estendida para outros modelos de cidades, o que é importante, considerando que diferentes locais podem ser melhor representados por outros modelos de Cidades Inteligentes. O MCS determinado neste trabalho representa o primeiro passo para fundamentar o conceito de um serviço de classificação com o objetivo de possibilitar o consumo de informação por diversos provedores de serviços urbanos e atores da cidade.

Esta pesquisa ajuda a trazer mais foco para a contribuição das redes sociais no processo de transformação das Cidades Inteligentes, conectando-as a modelos de indicadores conhecidos como ISO 37120. Muitas pesquisas têm se interessado em aplicar dados de RSO nas demandas da cidade, mas estão concentradas em domínios específicos, geralmente aplicando as técnicas de detecção de tópicos ou eventos, com numerosos trabalhos focando na localização da mensagem. Comunicações curtas e diretas nas redes sociais oriundas dos cidadãos,

como é o caso das reclamações às entidades relacionadas à prestação dos serviços públicos, não têm sido exploradas profundamente.

Acredita-se que um serviço de classificação, como o desenvolvido nesta pesquisa, conectado às dimensões da Cidade Inteligente pode contribuir significativamente no processo de melhoria das cidades, servindo como fonte de informações para vários usuários em potencial. Além disso, o trabalho descrito pode ser executado em outros contextos: outras redes sociais e outras localidades.

As principais contribuições deste trabalho são:

- um método detalhado para determinar os melhores MCS para classificar dados das redes sociais Twitter e Colab nas dimensões do padrão ISO 37120, e o próprio MCS selecionado;
- uma proposta de estrutura inovadora para um serviço de classificação, o ECMC, capaz de categorizar os dados de RSO em modelos conhecidos de Cidades Inteligentes e seus potenciais consumidores;
- uma base de conhecimento em português com termos relacionados a cada dimensão ISO.

Trabalhos futuros contemplam limitações descritas nesta dissertação, bem como a inclusão de outros padrões de Cidades Inteligentes (por exemplo, Modelo Europeu para Cidades Inteligentes, Iniciativa de Prosperidade das Cidades, atualizações da ISO); a avaliação de métodos adicionais para melhora do desempenho dos classificadores como explorar os parâmetros e técnicas de reamostragem na mineração de dados para balancear as classes; a aplicação de outros métodos de seleção de atributos e técnicas avançadas de processamento de texto, e avaliar a significância estatística na comparação dos métodos empregados. Além disso, será abordada a implementação do próprio serviço, testando os modelos gerados em um novo conjunto de dados em tempo real e a adição de análise de sentimentos sobre as postagens classificadas a fim de fornecer um ângulo diferente das impressões dos cidadãos sobre os serviços nas diferentes dimensões de Cidades Inteligentes.

REFERĘNCIAS

AGARWAL, B.; MITTAL, N. Text Classification Using Machine Learning Methods-A Survey. In: Babu B. et al. (Ed.). *Proceedings of the Second International Conference on Soft Computing for Problem Solving*. [S.l.]: Springer, New Delhi, 2014. p. 701–709.

AGGARWAL, C. C.; ZHAI, C. An introduction to text mining. In: AGGARWAL, C.; ZHAI, C. (Ed.). *Mining Text Data*. [S.l.]: Springer, Boston, MA, 2012. p. 1–10. ISBN 978-1-4614-3223-4.

AGGARWAL, C. C.; ZHAI, C. A survey of text classification algorithms. In: AGGARWAL, C.; ZHAI, C. (Ed.). *Mining Text Data*. [S.l.]: Springer, Boston, MA, 2012. p. 13–43. ISBN 978-1-4614-3223-4.

AGUILERA, U. et al. Citizen-centric data services for smarter cities. *Future Generation Computer Systems*, v. 76, p. 234–247, 2017.

ALLAHYARI, M. et al. A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques. In: *Proceedings of KDD Bigdas, Halifax, Canada*. [S.l.: s.n.], 2017. p. 13.

ANTHOPOULOS, L. Defining smart city architecture for sustainability. In: AL, E. T. et (Ed.). *Electronic Government and Electronic Participation*. [S.l.]: IOS Press, 2015. p. 140–147.

ANTHOPOULOS, L. G. *Understanding Smart Cities: A Tool for Smart Government or an Industrial Trick?* [S.l.]: Springer International Publishing, 2017. (Public Administration and Information Technology). ISBN 9783319570150.

BAEZA-YATES, R.; RIBEIRO-NETO, B. *Recuperação de Informação: Conceitos e Tecnologia das Máquinas de Busca*. [S.l.]: Bookman, Porto Alegre, 2013. ISBN 978-85-8260-049-8.

BELLO-ORGAZ, G.; HERNANDEZ-CASTRO, J.; CAMACHO, D. Detecting discussion communities on vaccination in twitter. *Future Generation Computer Systems*, v. 66, p. 125–136, 2017.

BELLO-ORGAZ, G.; J.JUNG, J.; CAMACHO, D. Social big data: Recent achievements and new challenges. *Information Fusion*, v. 28, p. 45–59, 2015.

BENCKE, L.; PEREZ, A. Análise dos principais modelos de indicadores para cidades sustentáveis e inteligentes. *Revista Nacional de Gerenciamento de Cidades*, v. 6, n. 37, 2018. <http://dx.doi.org/10.17271/2318847263720181754>.

BENCKE, L.; PEREZ, A.; ARMENDARIS, O. Rodovias inteligentes: uma visão geral sobre as tecnologias empregadas no brasil e no mundo. *iSys - Revista Brasileira de Sistemas de Informação*, v. 10, n. 4, p. 80–102, 2017. ISSN 1984-2902. <http://www.seer.unirio.br/index.php/isys/article/view/6609>.

BREIMAN, L. et al. *Classification and regression trees*. [S.l.]: CHAPMAN & HALUCRC, 1984. 358 p. ISBN: 0-412-04841-8.

BROWN, G. et al. Conditional Likelihood Maximisation: A Unifying Framework for Information Theoretic Feature Selection. *J. Mach. Learn. Res.*, JMLR.org, v. 13, p. 27–66, jan. 2012. ISSN 1532-4435. <http://dl.acm.org/citation.cfm?id=2188385.2188387>.

CARVALHO, J. P. et al. Misnis: An intelligent platform for twitter topic mining. *Expert Systems with Applications*, v. 89, p. 374–388, 2017.

CECHINEL, C.; CAMARGO, S. S. Mineração de dados educacionais: avaliação e interpretação de modelos de classificação. In: MAILLARD, P. A. J. et al. (Ed.). *Livro de Metodologia de Pesquisa em Informática na Educação. 1ed.* [S.l.]: (aceito para publicação), 2018. cap. 12.

CEREZO-COSTAS, H. et al. Discovering geo-dependent stories by combining density-based clustering and thread-based aggregation techniques. *Expert Systems with Applications*, v. 95, p. 32–42, 2018.

CHANDRASHEKAR, G.; SAHIN, F. A survey on feature selection methods. *Computers and Electrical Engineering*, v. 40, p. 16–28, 2013.

COCCHIA, A. Smart and digital city: A systematic literature review. In: DAMERI, R. P.; SABROUX, C. (Ed.). *How to create public and economic value with high technology in urban space*. [S.l.]: Springer, Cham, 2014, (Progress in IS). p. 13–43.

CONTERATTO, G. B. H. Algumas noções para aprimorar o tratamento de estruturas com predicado secundário. In: NOS, A. T. I.; PAIL, D. B. (Ed.). *Fundamentos linguísticos e computação*. [S.l.]: EDIPUCRS, 2015. ISBN 978-85-397-0661-7.

CORTES, C.; VAPNIK, V. Support-vector networks. *Machine Learning*, v. 20, n. 3, p. 273–297, Sep 1995. ISSN 1573-0565. <https://doi.org/10.1007/BF00994018>.

COSTA, K. A. da et al. Internet of Things: A survey on machine learning-based intrusion detection approaches. *Computer Networks*, v. 151, p. 147 – 157, 2019. ISSN 1389-1286. <http://www.sciencedirect.com/science/article/pii/S1389128618308739>.

DABIRI, S.; HEASLIP, K. Developing a twitter-based traffic event detection model using deep learning architectures. *Expert Systems with Applications*, v. 118, p. 425–439, 2018.

DAMERI, R. P. *Smart City Implementation - Creating Economic and Public Value in Innovative Urban Systems*. [S.l.]: Springer International Publishing, 2017. (Progress in Information Systems). ISBN 978-3-319-45766-6.

DAMERI, R. P.; COCHIA, A. Smart city and digital city: Twenty years of terminology evolution. *In X Conference of the Italian Chapter of AIS, ITAIS*, p. 1–8, 2013.

DOMÍNGUEZ, D. R. et al. Sensing the city with instagram: Clustering geolocated data for outlier detection. *Expert Systems with Applications*, v. 78, p. 319–333, 2017.

D'ANDREA, E. et al. Monitoring the public opinion about the vaccination topic from tweets analysis. *Expert Systems With Applications*, v. 116, p. 209–226, 2018.

ESPOSTE, A. de M. D. et al. Design and evaluation of a scalable smart city software platform with large-scale simulations. *Future Generation Computer Systems*, v. 93, p. 427–441, 2019.

FAN, R.-E. et al. LIBLINEAR: A Library for Large Linear Classification. *Journal of Machine Learning Research*, JMLR.org, v. 9, p. 1871–1874, jun. 2008. ISSN 1532-4435. <http://dl.acm.org/citation.cfm?id=1390681.1442794>.

FISCUS, J. G.; DODDINGTON, G. R. Topic Detection and Tracking Evaluation Overview. In: ALLAN, J. (Ed.). *Topic Detection and Tracking*. [S.l.]: Springer, Boston, MA, 2002. p. 17–31.

FLORES, F. N.; MOREIRA, V. P. Assessing the impact of stemming accuracy on information: A multilingual perspective. *Information Processing & Management*, v. 52, n. 5, p. 840–854, sep 2016.

FREUND, Y.; SCHAPIRE, R. E. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.*, Academic Press, Inc., Orlando, FL, USA, v. 55, n. 1, p. 119–139, ago. 1997. ISSN 0022-0000. <http://dx.doi.org/10.1006/jcss.1997.1504>.

GELERNTER, J.; MUSHEGIAN, N. Geo-parsing messages from microtext. *Transactions in GIS*, v. 15, n. 6, p. 753–773, 2011.

GUPTA, B. B. et al. Recent research in computational intelligence paradigms into security and privacy for online social networks (OSNs). *Future Generation Computer Systems*, v. 86, p. 851–854, 2018.

HAYKIN, S. *Neural Networks and Learning Machines - 3rd Edition*. [S.l.]: Pearson, 2009. 937 p.

HERRERA, F. et al. *Multiple Instance Learning - Foundations and Algorithms*. [S.l.]: Springer International Publishing, 2016. 241 p. ISBN: 978-3-319-47759-6.

HOLDEN, M. Sustainability indicator systems within urban governance: Usability analysis of sustainability indicator systems as boundary objects. *Ecological Indicators*, v. 32, p. 89–96, 2013.

HOQUE, N.; BHATTACHARYYA, D.; KALITA, J. A mutual information-based feature selection method. *Expert Systems with Applications*, v. 41, p. 6371–6385, 2014.

HUSSAIN, A.; CAMBRIA, E. Semi-supervised learning for big social data analysis. *Neurocomputing*, v. 275, p. 1662–1673, 2018.

IBGE. *Sinopse do Censo Demográfico 2010 do Brasil*. 2010. `https://censo2010.ibge.gov.br/sinopse/index.php?dados=8`. Online; accessed 06 January 2019.

ISO. *ISO 37120:2014*. 2014. `https://www.iso.org/obp/ui/#iso:std:iso:37120:ed-1:v1:en`. Online; accessed 06 January 2019.

JAMES, G. et al. *An Introduction to Statistical Learning: with Applications in R*. [S.l.]: Springer Science+Business Media, 2013. (Springer Texts in Statistics). ISBN 978-1-4614-7138-7.

JUNG, J. J. Recent advances on big data technologies and applications. *Mobile Networks and Applications*, v. 22, n. 4, p. 603–604, 2017.

KIM, J.; HASTAK, M. Social network analysis: Characteristics of online social networks after a disaster. *International Journal of Information Management*, v. 38, p. 86–96, 2018.

KIRITCHENKO, S.; ZHU, X.; MOHAMMAD, S. M. Sentiment analysis of short informal texts. *Journal of Artificial Intelligence Research*, v. 50, p. 723–762, 2014.

KOUSIOURIS, G. et al. An integrated information lifecycle management framework for exploiting social network data to identify dynamic large crowd concentration events in smart cities applications concentration events in smart cities applications. *Future Generation Computer Systems*, v. 78, p. 516–530, 2018.

KRASKOV, A.; STÖGBAUER, H.; GRASSBERGER, P. Estimating mutual information. *Phys. Rev. E*, American Physical Society, v. 69, p. 066138, Jun 2004.

KUMAR, M. et al. Feature selection and classification of microarray data using mapreduce based anova and k-nearest neighbor. *Procedia Computer Science*, v. 54, p. 301–310, 2015.

LABANI, M. et al. A novel multivariate filter method for feature selection in text classification problems. *Engineering Applications of Artificial Intelligence*, v. 70, p. 25–37, 2018.

LIU, C. et al. A new feature selection method based on a validity index of feature subset. *Pattern Recognition Letters*, v. 92, p. 1–8, 2017.

LIU, H.; SETIONO, R. Chi2: feature selection and discretization of numeric attributes. In: *Proceedings of 7th IEEE International Conference on Tools with Artificial Intelligence*. [S.l.: s.n.], 1995. p. 388–391. ISSN 1082-3409.

LIU, Y. et al. A statistical approach to participant selection in location-based social networks for offline event marketing. *Information Sciences*, v. 480, p. 90–108, 2019.

LIU, Y.; LOH, H. T.; SUN, A. Imbalanced text classification: A term weighting approach. *Expert Systems with Applications*, v. 36, n. 1, p. 690–701, 2009.

MANNING, C. D.; RAGHAVAN, P.; SCHüTZE, H. *An Introduction to Information Retrieval*. [S.l.]: Cambridge University Press, 2009. ISBN: 0521865719.

MIDDLETON, S. E.; MIDDLETON, L.; MODAFFERI, S. Real-Time Crisis Mapping of Natural Disasters Using Social Media. *IEEE Intelligent Systems*, v. 29, n. 2, p. 9–17, 2014.

MILIORIS, D. *Topic Detection and Classification in Social Networks - The Twitter Case*. [S.l.]: Springer International Publishing, 2018. (Communications Engineering, Networks). ISBN 9783319664149.

MITCHELL, T. M. *Machine Learning*. [S.l.]: McGraw-Hill Science/Engineering/Math, 1997. ISBN 0070428077.

MLADENIC, D. Feature selection in text mining. In: SAMMUT, C.; WEBB, G. I. (Ed.). *Encyclopedia of Machine Learning*. Boston, MA: Springer US, 2010. p. 406–410. ISBN 978-0-387-30164-8.

NOLASCO, D.; OLIVEIRA, J. Subevents detection through topic modeling in social media posts. *Future Generation Computer Systems*, v. 93, p. 290–303, 2019.

OLIVEIRA, M. G. a. de et al. A gold-standard social media corpus for urban issues. In: *Proceedings of the Symposium on Applied Computing*. New York, NY, USA: ACM, 2017. (SAC '17), p. 1011–1016. ISBN 978-1-4503-4486-9.

OLIVEIRA, M. G. de. *Ontology-driven Urban Issues Identification from Social Media*. Tese (Doutorado) — Federal University of Campina Grande, Center for Electrical and Computer Engineering, dez. 2016. <http://dspace.sti.ufcg.edu.br:8080/jspui/handle/riufcg/884>.

OSMAN, A. M. S. A novel big data analytics framework for smart cities. *Future Generation Computer Systems*, v. 91, p. 620–633, 2019.

PANAGIOTOU, N.; KATAKIS, I.; GUNOPULOS, D. Detecting Events in Online Social Networks: Definitions, Trends and Challenges. In: MICHAELIS, S.; PIATKOWSKI, N.; STOLPE, M. (Ed.). *Solving Large Scale Learning Tasks. Challenges and Algorithms*. [S.l.]: Springer, Cham, 2016. p. 42–84.

PANAGIOTOU, N. et al. Intelligent urban data monitoring for smart cities. *ECML PKDD 2016, Part III, Lecture Notes in Artificial Intelligence*, v. 9853, p. 177–192, 2016.

PEDREGOSA, F. et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, v. 12, p. 2825–2830, 2011.

PPGTIC. *Linhas de Pesquisa do Programa de Pós-Graduação em Tecnologias da Informação e Comunicação*. 2016. `http://ppgtic.ufsc.br/linhas-de-pesquisa/`. Online; acessado em 14 Fevereiro 2019.

PUIU, D. et al. Citypulse: Large scale data analytics framework for smart cities. *IEEE Access*, v. 4, p. 1086–1108, 2016. ISSN 2169-3536.

PURI, M. et al. Mapping Ordinances and Tweets Using Smart City Characteristics to Aid Opinion Mining. In: *Companion Proceedings of the The Web Conference 2018*. [S.l.]: International World Wide Web Conferences Steering Committee, 2018. (WWW '18), p. 1721–1728 . ISBN 978-1-4503-5640-4.

RENNIE, J. D. M. et al. Tackling the poor assumptions of naive bayes text classifiers. In: *Proceedings of the Twentieth International Conference on International Conference on Machine Learning*. AAAI Press, 2003. (ICML'03), p. 616–623. ISBN 1-57735-189-4. <http://dl.acm.org/citation.cfm?id=3041838.3041916>.

ROICK, O.; HEUSER, S. Location based social networks – definition, current state of the art and research agenda. *Transactions in GIS*, v. 17, n. 5, p. 763–784, 2013.

RUSSELL, S. J.; NORVIG, P. *Artificial Intelligence - A Modern Approach*. [S.l.]: Pearson, 2013. ISBN: 9781292153964.

SALTON, G.; WONG, A.; YANG, C. S. A vector space model for automatic indexing. *Commun. ACM*, ACM, New York, NY, USA, v. 18, n. 11, p. 613–620, nov. 1975. ISSN 0001-0782. <http://doi.acm.org/10.1145/361219.361220>.

SALTON, G.; YANG, C. S.; YU, C. T. A theory of term importance in automatic text analysis. *Journal of the Association for Information Science and Technology (JASIST)*, v. 26, n. 1, p. 33–44, fev. 1975.

SANTANA, E. F. Z. et al. Software platforms for smart cities: Concepts, requirements, challenges, and a unified reference architecture. *ACM Comput. Surv.*, ACM, New York, NY, USA, v. 50, n. 6, p. 78:1–78:37, nov. 2017. ISSN 0360-0300.

SANTOS, C. M. dos. *Classificação de Documentos com Processamento de Linguagem Natural*. 217 p. Dissertação (Mestrado) — Instituto Superior de Engenharia de Coimbra, Coimbra, 2015. <http://hdl.handle.net/10400.26/15293>.

SAPOUNTZI, A.; PSANNIS, K. Social networking data analysis tools & challenges. *Future Generation Computer Systems*, v. 86, p. 893–913, 2018.

SCIKIT-LEARN. *Supervised learning*. 2019. `https://scikit-learn.org`. Online; accessed 27 February 2019.

SEBASTIANI, F. Machine learning in automated text categorization. *ACM Computing Surveys (CSUR)*, v. 34, n. 1, p. 1–47, 2002.

SHARMA, P.; RAJPUT, S. Perspectives of smart cities: Introduction and overview. In: SHARMA, P.; RAJPUT, S. (Ed.). *Sustainable Smart Cities in India - Challenges and Future Perspectives*. [S.l.]: Springer International Publishing, 2017. p. 1–13.

SI, Y.; ZHANG, F.; LIU, W. An adaptive point-of-interest recommendation method for location-based social networks based on user activity and spatial features. *Knowledge-Based Systems*, v. 163, p. 267–282, 2019.

SILVA, B. C. D. da. *A face tecnológica dos estudos da linguagem: o processamento automático das línguas naturais*. Tese (Doutorado) — Universidade Estadual Paulista - UNESP, Faculdade de Ciências e Letras - Araraquara, 1996. <http://wiki.icmc.usp.br/images/a/ad/DiasDaSilva1996.pdf>.

SUZUKI, L. R. Smart Cities IoT: Enablers and Technology Road Map. *Smart Cities Networks - Through the Internet of Things, Springer Optimization and Its Applications (SOIA)*, v. 125, p. pp. 167–190, 2017.

THAKURIAH, P.; TILAHUN, N. Y.; ZELLNER, M. Introduction to Seeing Cities Through Big Data: Research, Methods and Applications in Urban Informatics. In: THAKURIAH, P.; TILAHUN, N.; ZELLNER, M. (Ed.). *Seeing Cities Through Big Data*. [S.l.]: Springer, Cham, 2017. p. 1–9.

TOKMAKOFF, A.; BILLINGTON, J. Consumer services in smart city Adelaide. In: BJERG, K.; BORREBY, K. (Ed.). *Proceedings of an International Cross-disciplinary Conference on Home-Oriented Informatics (HOIT)*. [S.l.]: University of Copenhagen, 1994. p. 201–210.

TUWIEN. *European Smart Cities Model*. 2015. `http://www.smart-cities.eu/`. Vienna University of Technology. Online; accessed 06 January 2019.

TWEEPY. *Documentation*. 2019. `https://https://tweepy.readthedocs.io/en/latest/`. Online; accessed 27 February 2019.

UNITED NATIONS. *World Urbanization Prospects*. 2018. `https://population.un.org/wup/`. Online; accessed 06 January 2019.

UYSAL, A. K.; GUNAL, S. The impact of preprocessing on text classification. *Information Processing & Management*, v. 50, n. 1, p. 104–112, 2014.

VLUYMANS, S. *Dealing with Imbalanced and Weakly Labelled Data in Machine Learning using Fuzzy and Rough Set Methods*. [S.l.]: Springer, 2019. (Studies in Computational Intelligence). ISBN 978-3-030-04663-7.

WANG, Q. et al. Classification of Private Tweets Using Tweet Content. In: *IEEE 11th International Conference on Semantic Computing (ICSC)*. [S.l.]: IEEE, 2017. p. 65–68.

WEILER, A.; GROSSNIKLAUS, M.; SCHOLL, M. H. Editorial: Survey and experimental analysis of event detection techniques for twitter. *The Computer Journal, Section C: Computational Intelligence, Machine Learning And Data Analytics*, v. 15, n. 6, p. 753–773, 2011.

WOJCIECH, C. Smart Governance for Smart Industries. In: *ICEGOV '13 Proceedings of the 7th International Conference on Theory and Practice of Electronic Governance*. [S.l.]: Seoul, Republic of Korea, 2013. p. 91–93.

ZDRAVESKI, V. et al. ISO-Standardized Smart City Platform Architecture and Dashboard. *IEEE Pervasive Computing*, v. 16, n. 2, p. 35–43, 2017.

ZHAI, C.; MASSUNG, S. *Text Data Management and Analysis - A Practical Introduction to Information Retrieval and Text Mining*. [S.l.]: ACM Books, 2016. ISBN: 978-1-97000-117-4.

ZHANG, H. The Optimality of Naive Bayes. In: *American Association for Artificial Intelligence*. [S.l.: s.n.], 2004.

ZHANG, Y. et al. HotML: A DSM-based machine learning system for social networks. *Journal of Computational Science*, v. 26, p. 478–487, 2018.

APÊNDICE A – Scores de Seleção e Pesos dos Atributos no MCS vencedor

Neste apêndice são apresentados mais detalhes a respeito da seleção de atributos efetuada nos dois melhores MCS da Tabela 7: para Twitter o experimento "S" e para a base Colab o experimento "I". Em ambos os experimentos a seleção de atributos ocorreu calculando-se a métrica da informação mútua (MI, *Mutual Information*), disponibilizada pela biblioteca Scikit-Learn. Também são apresentados os pesos utilizados pelo modelo vencedor Complement Naïve Bayes (CNB).

A.1 ATRIBUTOS DO TWITTER

Na base Twitter selecionou-se 4.688 atributos dos 31.688 originais. O resumo da distribuição dos scores MI calculados está na Tabela 8 e a lista dos 20 tokens com maior score MI encontra-se na Tabela 9.

Tabela 8 – Atributos do Twitter - Distribuição dos Scores MI

Métrica	**Todos os atributos (31688)**	**Atributos Selecionados (4688)**
Média	0,001111	0,003201
Desvio Padrão	0,001960	0,004421
Mínimo	0,000356	0,002244
1º Quartil (25%)	0,000361	0,002336
2º Quartil (50%)	0,000707	0,002447
3º Quartil (75%)	0,001501	0,002632
Máximo	0,212728	0,212728

Fonte: Elaborado pela autora.

Após o treinamento do classificador CNB a partir dos dados da matriz do Twitter $\boldsymbol{MT_{1950,4688}}$ contendo somente os atributos selecionados, foi possível verificar os pesos, baseados nos logaritmos naturais das probabilidades complementares, os quais foram calculados para cada par atributo/classe. Buscando por tokens que tivessem uma grande contribuição na classe calculou-se a diferença entre os dois maiores pesos dos atributos filtrando-se os casos onde esta diferença é maior que 20%. Desta forma foi possível visualizar atributos com um peso bem maior numa das classes, chamaremos estes atributos de "diferenciadores".

No conjunto de dados do Twitter estes atributos se concentram nas classes Água e Saneamento, Transportes, Energia e Águas Residuais, como pode ser visualizado nas Tabelas 10, 11 e 12. Esse resultado

Tabela 9 – Top 20 atributos selecionados por MI no Twitter

#	**Tokens**	**Score**
1	agua	0,212727823
2	luz	0,140057923
3	transit	0,04497428
4	esgot	0,041414212
5	onibus	0,035561102
6	rua	0,034220655
7	bairr	0,028114791
8	falt luz	0,023928332
9	falt	0,023720887
10	falt agua	0,023701084
11	morador	0,023646906
12	previsa	0,023303582
13	burac	0,023232905
14	eptc	0,022047349
15	energ	0,021530572
16	morador rua	0,0210185
17	ano	0,020318372
18	contat	0,020042745
19	carr	0,019464383
20	estacion	0,01917976

Fonte: Elaborado pela autora.

corresponde às classes de melhor F1-score (Figura 35 da Seção 5.1). As medidas estatísticas dos pesos de todos os 4.688 atributos usados no treinamento de CNB podem ser visualizadas para cada classe na Tabela 13. Nas Tabelas 14, 15 e 16 são apresentados 20 dos atributos com maior peso em cada classe.

A.2 ATRIBUTOS DO COLAB

Na base Colab selecionou-se 99.340 atributos dos 1.089.340 originais. O resumo da distribuição dos scores MI calculados está na Tabela 17 e a lista dos 20 tokens com maior score MI encontra-se na Tabela 18.

Após o treinamento do classificador CNB a partir dos dados da matriz do Colab $\boldsymbol{MC_{65066,99340}}$ contendo somente os atributos selecionados, foi possível verificar os pesos, baseados nos logaritmos naturais das probabilidades complementares, os quais foram calculados para cada par atributo/classe. Buscando por tokens que tivessem uma grande contribuição na classe calculou-se a diferença entre os dois maiores pesos dos atributos filtrando-se os casos onde esta diferença é maior que 30%. Optou-se por um percentual maior que o Twitter pois o volume de atributos selecionados para a base Colab é bem maior: usando 20% temos 1026 atributos o que não é viável representar neste documento.

No conjunto de dados do Colab os atributos diferenciadores, usando o limite de 30%, se concentram nas classes Transportes e Energia, como pode ser visualizado nas Tabelas 19, 20, 21 e 22. Esse resultado corresponde às classes de melhor F1-score (Figura 35 da Seção 5.1). As medidas estatísticas dos pesos de todos os 99.340 atributos usados no treinamento de CNB podem ser visualizadas para cada classe na Tabela 13. Nas Tabelas 24 e 25 estão apresentados 20 dos atributos com maior peso em cada classe.

Tabela 10 – Pesos de Atributos Diferenciadores - Twitter (cont.)

	Água e Saneament	Transportes	Energia	Nenhum	Recreação	Resíduos Sólidos	Economia	Finanças	Águas Residuais	Segurança	Educação	Meio Ambiente	Saúde	Telecomunicações	Habitação	Fogo e Resp Desastre	% acima da 2a classe
tokens \ classes	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	
agua	7,63	4,72	4,85	4,69	4,88	4,88	4,89	4,88	4,89	4,89	4,89	4,88	4,87	4,90	4,89	4,89	56%
transit	5,69	8,39	5,71	5,58	5,73	5,73	5,74	5,73	5,74	5,74	5,74	5,73	5,71	5,75	5,74	5,73	46%
luz	6,00	5,89	6,02	8,69	6,04	6,05	6,06	6,05	6,06	6,06	6,06	6,05	6,03	6,07	6,06	6,05	43%
faix	6,00	5,89	6,02	8,69	6,04	6,05	6,06	6,05	6,06	6,06	6,06	6,05	6,03	6,07	6,06	6,05	43%
castel	7,10	9,78	7,12	6,96	7,14	7,15	7,15	7,15	7,16	7,16	7,16	7,15	7,13	7,17	7,15	7,15	36%
castel branc	6,57	9,09	6,59	6,46	6,61	6,61	6,62	6,61	6,62	6,62	6,63	6,62	6,59	6,63	6,62	6,61	37%
falt agua	7,30	9,78	7,31	7,15	7,34	7,34	7,35	7,34	7,35	7,35	7,35	7,34	7,32	7,36	7,35	7,34	33%
onibus	5,34	5,18	7,76	5,21	5,38	5,39	5,39	5,41	5,39	5,40	5,41	5,39	5,37	5,41	5,39	5,41	43%
pedestr	7,37	9,78	7,39	7,22	7,41	7,42	7,42	7,42	7,42	7,43	7,43	7,42	7,39	7,44	7,42	7,42	31%
semafor	9,93	7,38	7,56	7,39	7,58	7,58	7,59	7,58	7,59	7,59	7,60	7,58	7,56	7,60	7,59	7,58	31%
estacion	7,45	7,29	7,47	9,79	7,49	7,50	7,50	7,50	7,50	7,51	7,51	7,50	7,47	7,52	7,50	7,50	30%
sinaleir	7,45	7,29	7,47	9,79	7,49	7,50	7,50	7,50	7,50	7,51	7,51	7,50	7,47	7,52	7,50	7,50	30%
branc	5,93	5,81	5,93	8,00	5,95	5,96	5,96	5,97	5,96	5,97	5,97	5,96	5,93	5,98	5,96	5,95	34%
sinaliz	7,30	9,78	7,31	7,15	7,34	7,34	7,35	7,34	7,35	7,35	7,35	7,34	7,32	7,36	7,35	7,34	33%
burac	6,15	8,17	6,15	6,01	6,17	6,17	6,18	6,20	6,18	6,18	6,19	6,18	6,15	6,19	6,20	6,17	32%
vias	7,37	9,78	7,39	7,22	7,41	7,42	7,42	7,42	7,42	7,43	7,43	7,42	7,39	7,44	7,42	7,42	31%
retorn agua	9,93	7,38	7,56	7,39	7,58	7,58	7,59	7,58	7,59	7,59	7,60	7,58	7,56	7,60	7,59	7,58	31%
av castel	7,45	7,29	7,47	9,79	7,49	7,50	7,50	7,50	7,50	7,51	7,51	7,50	7,47	7,52	7,50	7,50	30%
av castel branc	7,45	7,29	7,47	9,79	7,49	7,50	7,50	7,50	7,50	7,51	7,51	7,50	7,47	7,52	7,50	7,50	30%
bolsonar	7,45	7,29	7,47	9,79	7,49	7,50	7,50	7,50	7,50	7,51	7,51	7,50	7,47	7,52	7,50	7,50	30%
psol	6,94	6,84	6,96	9,10	6,98	6,99	6,99	6,99	6,99	7,00	7,00	6,99	6,96	7,01	6,99	6,98	30%
transport	6,94	9,09	6,96	6,84	6,98	6,99	6,99	6,99	6,99	7,00	7,00	6,99	6,96	7,01	6,99	6,98	30%
ciclov	7,54	9,78	7,56	7,39	7,58	7,58	7,59	7,58	7,59	7,59	7,60	7,58	7,56	7,60	7,59	7,58	29%
transport public	7,54	9,78	7,56	7,39	7,58	7,58	7,59	7,58	7,59	7,59	7,60	7,58	7,56	7,60	7,59	7,58	29%
esgot	7,16	7,07	7,18	7,02	7,20	7,21	7,21	7,21	9,30	7,22	7,22	7,21	7,19	7,23	7,21	7,21	29%
agua nov	9,93	7,58	7,76	7,59	7,78	7,78	7,79	7,78	7,79	7,79	7,80	7,79	7,76	7,80	7,79	7,78	27%
emporcalh	7,63	7,48	7,65	9,79	7,67	7,68	7,68	7,68	7,69	7,69	7,69	7,68	7,66	7,70	7,68	7,68	27%
alca	7,63	9,78	7,65	7,49	7,67	7,68	7,68	7,68	7,69	7,69	7,69	7,68	7,66	7,70	7,68	7,68	27%
dmae	9,24	7,07	7,25	7,15	7,27	7,27	7,28	7,27	7,28	7,28	7,29	7,27	7,25	7,29	7,28	7,27	27%
aguard contat	7,85	7,70	7,87	7,71	7,90	7,90	7,91	7,90	9,99	7,91	7,91	7,90	7,88	7,92	7,91	7,90	26%
aguard contat referent	7,85	7,70	7,87	7,71	7,90	7,90	7,91	7,90	9,99	7,91	7,91	7,90	7,88	7,92	7,91	7,90	26%
ano meio	7,85	7,70	7,87	7,71	7,90	7,90	7,91	7,90	9,99	7,91	7,91	7,90	7,88	7,92	7,91	7,90	26%
ano meio obrig	7,85	7,70	7,87	7,71	7,90	7,90	7,91	7,90	9,99	7,91	7,91	7,90	7,88	7,92	7,91	7,90	26%
contat referent	7,85	7,70	7,87	7,71	7,90	7,90	7,91	7,90	9,99	7,91	7,91	7,90	7,88	7,92	7,91	7,90	26%

Fonte: Elaborado pela Autora.

Tabela 11 – Pesos de Atributos Diferenciadores - Twitter (cont.)

	Água e Saneament	Transportes	Energia	Nenhum	Recreação	Residuos Sólidos	Economia	Finanças	Aguas Residuais	Segurança	Educação	Meio Ambiente	Saúde	Telecomunicações	Habitação	Fogo e Resp Desastre	
tokens \ classes	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	% acima da 2a classe
contat referent esgot	7,85	7,70	7,87	7,71	7,90	7,90	7,91	7,90	9,99	7,91	7,91	7,90	7,88	7,92	7,91	7,90	26%
esgot pati	7,85	7,70	7,87	7,71	7,90	7,90	7,91	7,90	9,99	7,91	7,91	7,90	7,88	7,92	7,91	7,90	26%
esgot pati ano	7,85	7,70	7,87	7,71	7,90	7,90	7,91	7,90	9,99	7,91	7,91	7,90	7,88	7,92	7,91	7,90	26%
meio obrig	7,85	7,70	7,87	7,71	7,90	7,90	7,91	7,90	9,99	7,91	7,91	7,90	7,88	7,92	7,91	7,90	26%
pati	7,85	7,70	7,87	7,71	7,90	7,90	7,91	7,90	9,99	7,91	7,91	7,90	7,88	7,92	7,91	7,90	26%
pati ano	7,85	7,70	7,87	7,71	7,90	7,90	7,91	7,90	9,99	7,91	7,91	7,90	7,88	7,92	7,91	7,90	26%
pati ano meio	7,85	7,70	7,87	7,71	7,90	7,90	7,91	7,90	9,99	7,91	7,91	7,90	7,88	7,92	7,91	7,90	26%
referent esgot	7,85	7,70	7,87	7,71	7,90	7,90	7,91	7,90	9,99	7,91	7,91	7,90	7,88	7,92	7,91	7,90	26%
referent esgot pati	7,85	7,70	7,87	7,71	7,90	7,90	7,91	7,90	9,99	7,91	7,91	7,90	7,88	7,92	7,91	7,90	26%
respond telefon aguard	7,85	7,70	7,87	7,71	7,90	7,90	7,91	7,90	9,99	7,91	7,91	7,90	7,88	7,92	7,91	7,90	26%
telefon aguard	7,85	7,70	7,87	7,71	7,90	7,90	7,91	7,90	9,99	7,91	7,91	7,90	7,88	7,92	7,91	7,90	26%
telefon aguard (	7,85	7,70	7,87	7,71	7,90	7,90	7,91	7,90	9,99	7,91	7,91	7,90	7,88	7,92	7,91	7,90	26%
alca acess	7,74	9,78	7,76	7,59	7,78	7,78	7,79	7,78	7,79	7,79	7,80	7,79	7,76	7,80	7,79	7,78	25%
crat	7,74	9,78	7,76	7,59	7,78	7,78	7,79	7,78	7,79	7,79	7,80	7,79	7,76	7,80	7,79	7,78	25%
pass livr	7,74	9,78	7,76	7,59	7,78	7,78	7,79	7,78	7,79	7,79	7,80	7,79	7,76	7,80	7,79	7,78	25%
agua bairr	9,93	7,70	7,87	7,71	7,90	7,90	7,91	7,90	7,91	7,91	7,91	7,90	7,88	7,92	7,91	7,90	25%
desd ontem	9,93	7,70	7,87	7,71	7,90	7,90	7,91	7,90	7,91	7,91	7,91	7,90	7,88	7,92	7,91	7,90	25%
volt agua	9,93	7,70	7,87	7,71	7,90	7,90	7,91	7,90	7,91	7,91	7,91	7,90	7,88	7,92	7,91	7,90	25%
acident	6,20	7,83	6,22	6,21	6,24	6,24	6,25	6,24	6,25	6,25	6,26	6,24	6,22	6,26	6,25	6,24	25%
tirotei	7,99	7,83	8,01	7,84	8,03	8,04	8,04	8,04	8,04	9,99	8,05	8,04	8,01	8,06	8,04	8,03	24%
falt luz	7,04	6,89	8,86	6,90	7,09	7,09	7,10	7,15	7,10	7,10	7,10	7,09	7,07	7,11	7,10	7,15	24%
bent	7,23	9,09	7,25	7,08	7,27	7,27	7,28	7,27	7,28	7,28	7,29	7,34	7,25	7,29	7,28	7,27	24%
iptu	7,99	7,83	8,01	7,84	8,03	8,04	8,04	9,98	8,04	8,05	8,05	8,04	8,01	8,06	8,04	8,03	24%
vacin	7,99	7,83	8,01	7,84	8,03	8,04	8,04	8,04	8,04	8,05	8,05	8,04	9,96	8,06	8,04	8,03	24%
cruzament	7,85	9,78	7,87	7,71	7,90	7,90	7,91	7,90	7,91	7,91	7,91	7,90	7,88	7,92	7,91	7,90	23%
flux	7,85	9,78	7,87	7,71	7,90	7,90	7,91	7,90	7,91	7,91	7,91	7,90	7,88	7,92	7,91	7,90	23%
manobr	7,85	9,78	7,87	7,71	7,90	7,90	7,91	7,90	7,91	7,91	7,91	7,90	7,88	7,92	7,91	7,90	23%
par onibus	7,85	9,78	7,87	7,71	7,90	7,90	7,91	7,90	7,91	7,91	7,91	7,90	7,88	7,92	7,91	7,90	23%
luz desd	7,99	7,83	9,95	7,84	8,03	8,04	8,04	8,04	8,04	8,05	8,05	8,04	8,01	8,06	8,04	8,03	23%
abastec	8,84	6,95	7,12	7,08	7,14	7,15	7,15	7,15	7,16	7,16	7,16	7,15	7,13	7,17	7,15	7,15	23%
agua desd	9,93	7,83	8,01	7,84	8,03	8,04	8,04	8,04	8,04	8,05	8,05	8,04	8,01	8,06	8,04	8,03	23%
previsa retorn a	9,93	7,83	8,01	7,84	8,03	8,04	8,04	8,04	8,04	8,05	8,05	8,04	8,01	8,06	8,04	8,03	23%
eptc	5,93	7,38	5,95	5,96	5,97	5,97	6,00	5,97	5,98	5,98	5,99	5,97	5,95	5,99	5,98	5,97	23%

Fonte: Elaborado pela Autora.

Tabela 12 – Pesos de Atributos Diferenciadores - Twitter (conclusão)

	Água e Saneament	Transportes	Energia	Nenhum	Recreação	Resíduos Sólidos	Economia	Finanças	Águas Residuais	Segurança	Educação	Meio Ambiente	Saúde	Telecomunicações	Habitação	Fogo e Resp Desastre	% acima da 2a classe
tokens \ classes	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	
flanelinh	6,99	8,68	7,01	6,84	7,09	7,04	7,04	7,04	7,05	7,10	7,05	7,04	7,02	7,06	7,04	7,04	22%
nome	6,80	6,69	6,82	8,40	6,84	6,85	6,85	6,85	6,85	6,86	6,86	6,89	6,87	6,87	6,85	6,84	22%
energ	7,04	6,78	8,57	6,79	6,98	6,99	6,99	6,99	6,99	7,00	7,00	6,99	6,96	7,01	6,99	7,04	22%
bust	8,14	7,99	8,16	8,00	9,98	8,19	8,20	8,19	8,20	8,20	8,20	8,19	8,17	8,21	8,20	8,19	22%
cear	7,99	9,78	8,01	7,84	8,03	8,04	8,04	8,04	8,04	8,05	8,05	8,04	8,01	8,06	8,04	8,03	21%
ciclist	7,99	9,78	8,01	7,84	8,03	8,04	8,04	8,04	8,04	8,05	8,05	8,04	8,01	8,06	8,04	8,03	21%
congestion	7,99	9,78	8,01	7,84	8,03	8,04	8,04	8,04	8,04	8,05	8,05	8,04	8,01	8,06	8,04	8,03	21%
grav	7,99	9,78	8,01	7,84	8,03	8,04	8,04	8,04	8,04	8,05	8,05	8,04	8,01	8,06	8,04	8,03	21%
par carl	7,99	9,78	8,01	7,84	8,03	8,04	8,04	8,04	8,04	8,05	8,05	8,04	8,01	8,06	8,04	8,03	21%
par carl gom	7,99	9,78	8,01	7,84	8,03	8,04	8,04	8,04	8,04	8,05	8,05	8,04	8,01	8,06	8,04	8,03	21%
pont guaib	7,99	9,78	8,01	7,84	8,03	8,04	8,04	8,04	8,04	8,05	8,05	8,04	8,01	8,06	8,04	8,03	21%
rotul	7,99	9,78	8,01	7,84	8,03	8,04	8,04	8,04	8,04	8,05	8,05	8,04	8,01	8,06	8,04	8,03	21%
sent centr	7,99	9,78	8,01	7,84	8,03	8,04	8,04	8,04	8,04	8,05	8,05	8,04	8,01	8,06	8,04	8,03	21%
tap burac	7,99	9,78	8,01	7,84	8,03	8,04	8,04	8,04	8,04	8,05	8,05	8,04	8,01	8,06	8,04	8,03	21%
travess	7,99	9,78	8,01	7,84	8,03	8,04	8,04	8,04	8,04	8,05	8,05	8,04	8,01	8,06	8,04	8,03	21%
energ eletr	8,14	7,99	9,95	8,00	8,18	8,19	8,20	8,19	8,20	8,20	8,20	8,19	8,17	8,21	8,20	8,19	21%
hor luz	8,14	7,99	9,95	8,00	8,18	8,19	8,20	8,19	8,20	8,20	8,20	8,19	8,17	8,21	8,20	8,19	21%
agua aqui	9,93	7,99	8,16	8,00	8,18	8,19	8,20	8,19	8,20	8,20	8,20	8,19	8,17	8,21	8,20	8,19	21%
belem velh agu	9,93	7,99	8,16	8,00	8,18	8,19	8,20	8,19	8,20	8,20	8,20	8,19	8,17	8,21	8,20	8,19	21%
fic agua	9,93	7,99	8,16	8,00	8,18	8,19	8,20	8,19	8,20	8,20	8,20	8,19	8,17	8,21	8,20	8,19	21%
torneir	9,93	7,99	8,16	8,00	8,18	8,19	8,20	8,19	8,20	8,20	8,20	8,19	8,17	8,21	8,20	8,19	21%
velh agua	9,93	7,99	8,16	8,00	8,18	8,19	8,20	8,19	8,20	8,20	8,20	8,19	8,17	8,21	8,20	8,19	21%
livr	7,45	9,09	7,47	7,39	7,49	7,50	7,50	7,50	7,50	7,51	7,51	7,50	7,47	7,52	7,50	7,50	21%
beir	7,54	9,09	7,56	7,49	7,58	7,58	7,59	7,58	7,59	7,59	7,60	7,58	7,56	7,60	7,59	7,58	20%
beir rio	7,54	9,09	7,56	7,49	7,58	7,58	7,59	7,58	7,59	7,59	7,60	7,58	7,56	7,60	7,59	7,58	20%
cobertur	7,54	9,09	7,56	7,49	7,58	7,58	7,59	7,58	7,59	7,59	7,60	7,58	7,56	7,60	7,59	7,58	20%
tabel	7,54	9,09	7,56	7,49	7,58	7,58	7,59	7,58	7,59	7,59	7,60	7,58	7,56	7,60	7,59	7,58	20%
tap	7,54	9,09	7,56	7,49	7,58	7,58	7,59	7,58	7,59	7,59	7,60	7,58	7,56	7,60	7,59	7,58	20%
belem velh	8,55	6,89	7,12	6,96	7,09	7,15	7,10	7,09	7,10	7,10	7,10	7,09	7,07	7,11	7,10	7,09	20%
vot	7,16	7,01	7,18	8,69	7,20	7,27	7,21	7,27	7,22	7,22	7,22	7,21	7,19	7,23	7,21	7,21	20%

Fonte: Elaborado pela Autora.

Tabela 13 – Distribuição dos Pesos nas Classes - Twitter

	0	1	2	3	4
count	4688.000000	4688.000000	4688.000000	4688.000000	4688.000000
mean	8.923601	8.815437	8.931524	8.797514	8.999073
std	0.732970	0.636308	0.738006	0.629415	0.810637
min	5.082180	4.717255	4.854124	4.689388	4.875709
25%	9.241064	9.086703	9.260843	8.690642	8.876963
50%	9.241064	9.086703	9.260843	9.096107	9.282428
75%	9.241064	9.086703	9.260843	9.096107	9.282428
max	9.934211	9.779850	9.953991	9.789254	9.975576

	5	6	7	8	9
count	4688.000000	4688.000000	4688.000000	4688.000000	4688.000000
mean	8.967412	8.976962	8.985345	8.956036	8.944460
std	0.774423	0.784632	0.795165	0.753142	0.743029
min	4.881045	4.887595	4.882016	4.890028	4.891174
25%	8.882299	8.888849	8.883270	8.891282	8.892428
50%	9.287764	9.294314	9.288736	9.296747	9.297893
75%	9.287764	9.294314	9.288736	9.296747	9.297893
max	9.980911	9.987461	9.981883	9.989895	9.991040

	10	11	12	13	14
count	4688.000000	4688.000000	4688.000000	4688.000000	4688.000000
mean	8.966798	8.983329	9.012734	8.957020	8.972803
std	0.772018	0.792120	0.826246	0.758760	0.780388
min	4.893508	4.882432	4.866070	4.901156	4.887503
25%	8.894762	8.883686	8.861208	8.902410	8.888757
50%	9.300227	9.289152	9.266673	9.307875	9.294222
75%	9.300227	9.289152	9.266673	9.307875	9.294222
max	9.993374	9.982299	9.959821	10.001023	9.987369

	15
count	4688.000000
mean	8.983117
std	0.791313
min	4.892667
25%	8.881651
50%	9.287116
75%	9.287116
max	9.980263

LENGEDA:
- **count:** número de tokens
- **mean:** média
- **std:** desvio padrão
- **min:** menor peso
- **25%:** 1º quartil
- **50%:** 2º quartil
- **75%:** 3º quartil

0 - Água e Saneamento
1 - Transportes
2 - Energia
3 - Nenhum
4 - Recreação
5 - Resíduos Sólidos
6 - Economia
7 - Finanças
8 - Águas Residuais
9 - Segurança
10 - Educação
11 - Meio Ambiente
12 - Saúde
13 - Telecomunicações
14 - Habitação
15 - Fogo e Resp Desastre

Fonte: Elaborado pela Autora.

Tabela 14 – Atributos e Pesos por Classe - Twitter (cont.)

Água e Saneamento		Transportes		Energia	
0		1		2	
9,9342	agua abert	9,7798	acessibil	9,9540	dias ilumin
9,9342	agua abert morr	9,7798	ruas port alegr	9,9540	dias ilumin public
9,9342	falt agua	9,7798	av sertori	9,9540	dias nada resolv
9,9342	agua algum previsa	9,7798	bonifaci	9,9540	energ resident
9,9342	agua bairr lag	9,7798	cobertur par	9,9540	falt luz aqui
9,9342	agua abert	9,7798	cobertur par carl	9,9540	falt luz bairr
9,9342	agua abert morr	9,7798	sent centr bairr	9,9540	geladeir
9,9342	agua aind	9,7798	trafeg	9,9540	hor energ eletr
9,9342	agua algum previsa	9,7798	flanelinh remov	9,9540	ilumin public funcion
9,9342	agua bairr lag	9,7798	interior	9,9540	liga luz
9,9342	nova algum previsa	9,7798	jose bonifaci	9,9540	luz agor
9,9342	previsa hor	9,7798	loureir	9,9540	luz ja
9,9342	previsa retorn abastec	9,7798	minut esper	9,9540	qued luz
9,9342	potavel	9,7798	nilo pecanh	9,9540	sobr falt luz
9,9342	retorn abastec	9,7798	passageir	9,9540	tao brincadeir
9,9342	vila nova agua	9,7798	pecanh	9,9540	trint dias ilumin
9,9342	agua aqui	9,7798	proib estacion	9,9540	meia fase
9,9342	torneir	9,7798	rad	9,9540	ilumin public
9,9342	desd ontem	9,7798	rotul cui	9,9540	energ eletr
9,9342	retorn agua	9,7798	ruas port	9,9540	hor luz
Nenhum		Recreação		Resíduos Sólidos	
3		4		5	
9,7893	emporcalh	9,9756	acontec durant dia	9,9809	pendur saco lixo
9,7893	bolsonar	9,9756	apresent tres film	9,9809	popul porc
9,7893	acredit tod saim	9,9756	banc descent quiosqu	9,9809	poss descart forn
9,7893	angu nao fez	9,9756	banheir	9,9809	prec paga jog
9,0961	bastant inform	9,9756	brinqu redenca	9,9809	reclam cidad suja
9,0961	cabe investig	9,9756	catamara	9,9809	recolh lixo
9,0961	funcion mau	9,9756	cinematec	9,9809	relaca aleg limpez
9,0961	implement	9,9756	cultur cidad	9,9809	retalh pendur
9,0961	porqu site rodoviar	9,9756	descaracteriz prac matriz	9,9809	saco lixo
9,0961	prefeitur consig acess	9,9756	esport sim	9,9809	selet aqui
9,0961	ai popul cam	9,9756	fubangueir	9,9809	separ lixo
9,0961	aind bem tod	9,9756	guaib pra elit	9,9809	suja cidad aind
9,0961	bat professor	9,9756	obra arte	9,9809	ter problem colet
9,0961	gent envolv	9,9756	orla guaib bonit	9,9809	tudo entulh
9,0961	guerreir marchezan	9,9756	patrimoni public	9,9809	ultim reclam
9,0961	guerreir marchezan vergonh	9,9756	program ingress compr	9,9809	vez caminha colet
9,0961	hoje absurd	9,9756	quadr tenn vol	9,9809	lixeir
9,0961	invasa ocupaca	9,9756	usufru	9,9809	lixo seco
9,0961	marchezan vergonh	9,9756	vol esport	9,9809	vila ecolog
9,0961	orgulh	9,9756	zero musical	9,9809	dmlu

Fonte: Elaborado pela Autora.

Tabela 15 – Atributos e Pesos por Classe - Twitter (cont.)

Economia		Finanças		Águas Residuais	
6		7		8	
9,9875	ano despes	9,9819	ajust financeir	9,9899	alert esgot
9,9875	aument salari	9,9819	aposent pag	9,9899	doming tod
9,9875	cest basic	9,9819	dinheir ipva	9,9899	esgot invad
9,9875	codig consumidor	9,9819	fic iptu	9,9899	esgot transbord
9,9875	compr agor	9,9819	financ colet pag	9,9899	esper lig
9,9875	desempreg repass	9,9819	fiscal pertenc	9,9899	esper lig esgot
9,9875	despes	9,9819	fortun paga	9,9899	frent ceee joaquim
9,9875	diesel	9,9819	funcion maquin public	9,9899	instal esgot invad
9,9875	empreendedor	9,9819	itbi	9,9899	lig nov
9,9875	empresari	9,9819	mult nojo	9,9899	pass contat
9,9875	falt vergonh	9,9819	nojo dess gent	9,9899	retorn instal
9,9875	gasolin cara	9,9819	nota fiscal	9,9899	transbord
9,9875	litr gasolin	9,9819	pertenc impost recolh	9,9899	esgot espalh
9,9875	mesm acredit	9,9819	planej	9,9899	pois caix quebr
9,9875	precis volt creciment	9,9819	popul invist saud	9,9899	quebr pag impost
9,9875	repass	9,9819	quant vai fic	9,9899	esgot pati
9,9875	sab gasolin	9,9819	vereador vot	9,9899	referent esgot
9,9875	tripl prec	9,9819	estar defas	9,9899	respond telefon aguard
9,9875	venc solicit	9,9819	financ	9,9899	telefon aguard
9,9875	volt creciment	9,9819	fortun	9,9899	telefon aguard contat
Segurança		**Educação**		**Meio Ambiente**	
9		10		11	
9,9910	ar vari	9,9934	alun	9,9823	ambient
9,9910	cham brig	9,9934	bat professor	9,9823	arvor inteir
9,9910	jov roub	9,9934	clar professor	9,9823	arvor podr
9,9910	jov roub dois	9,9934	educ reflex	9,9823	arvor ruas
9,9910	legal jov	9,9934	encerr hoje	9,9823	barulh
9,9910	legal jov roub	9,9934	erro mec	9,9823	barulh elev
9,9910	mata pesso	9,9934	escol atend	9,9823	bomb jardim
9,9910	pesso assalt	9,9934	fal ingles	9,9823	botan ruid
9,9910	roub dois	9,9934	faz remanej	9,9823	clareir
9,9910	roub dois dias	9,9934	irao encerr	9,9823	cort mato
9,9910	susep	9,9934	muit escol	9,9823	equip manutenca
9,9910	tirotei helicopter	9,9934	ocupaca	9,9823	faco denunc poluica
9,9910	mata	9,9934	pro brasil	9,9823	imprens
9,9910	tirotei	9,9934	professor jog	9,9823	meio ambient
9,2979	acontec durant dia	9,9934	professor setor public	9,9823	pass atenca
9,2979	anos conhec	9,9934	site chei	9,9823	pluviometr
9,2979	cintur pra cima	9,9934	sociedad educ	9,9823	recuper arroi nome
9,2979	continu la desloc	9,9934	superior	9,9823	ruid alto
9,2979	cada gesta	9,9934	vergonh pro brasil	9,9823	sonor
9,2979	diz nao	9,9934	mec	9,9823	suj meio ambient

Fonte: Elaborado pela Autora.

Tabela 16 – Atributos e Pesos por Classe - Twitter (conclusão)

Economia		Finanças		Águas Residuais	
6		7		8	
9,9875	ano despes	9,9819	ajust financeir	9,9899	alert esgot
9,9875	aument salari	9,9819	aposent pag	9,9899	doming tod
9,9875	cest basic	9,9819	dinheir ipva	9,9899	esgot invad
9,9875	codig consumidor	9,9819	fic iptu	9,9899	esgot transbord
9,9875	compr agor	9,9819	financ colet pag	9,9899	esper lig
9,9875	desempreg repass	9,9819	fiscal pertenc	9,9899	esper lig esgot
9,9875	despes	9,9819	fortun paga	9,9899	frent ceee joaquim
9,9875	diesel	9,9819	funcion maquin public	9,9899	instal esgot invad
9,9875	empreendedor	9,9819	itbi	9,9899	lig nov
9,9875	empresari	9,9819	mult nojo	9,9899	pass contat
9,9875	falt vergonh	9,9819	nojo dess gent	9,9899	retorn instal
9,9875	gasolin cara	9,9819	nota fiscal	9,9899	transbord
9,9875	litr gasolin	9,9819	pertenc impost recolh	9,9899	esgot espalh
9,9875	mesm acredit	9,9819	planej	9,9899	pois caix quebr
9,9875	precis volt creciment	9,9819	popul invist saud	9,9899	quebr pag impost
9,9875	repass	9,9819	quant vai fic	9,9899	esgot pati
9,9875	sab gasolin	9,9819	vereador vot	9,9899	referent esgot
9,9875	tripl prec	9,9819	estar defas	9,9899	respond telefon aguard
9,9875	venc solicit	9,9819	financ	9,9899	telefon aguard
9,9875	volt creciment	9,9819	fortun	9,9899	telefon aguard contat
Segurança		**Educação**		**Meio Ambiente**	
9		10		11	
9,9910	ar vari	9,9934	alun	9,9823	ambient
9,9910	cham brig	9,9934	bat professor	9,9823	arvor inteir
9,9910	jov roub	9,9934	clar professor	9,9823	arvor podr
9,9910	jov roub dois	9,9934	educ reflex	9,9823	arvor ruas
9,9910	legal jov	9,9934	encerr hoje	9,9823	barulh
9,9910	legal jov roub	9,9934	erro mec	9,9823	barulh elev
9,9910	mata pesso	9,9934	escol atend	9,9823	bomb jardim
9,9910	pesso assalt	9,9934	fal ingles	9,9823	botan ruid
9,9910	roub dois	9,9934	faz remanej	9,9823	clareir
9,9910	roub dois dias	9,9934	irao encerr	9,9823	cort mato
9,9910	susep	9,9934	muit escol	9,9823	equip manutenca
9,9910	tirotei helicopter	9,9934	ocupaca	9,9823	faco denunc poluica
9,9910	mata	9,9934	pro brasil	9,9823	imprens
9,9910	tirotei	9,9934	professor jog	9,9823	meio ambient
9,2979	acontec durant dia	9,9934	professor setor public	9,9823	pass atenca
9,2979	anos conhec	9,9934	site chei	9,9823	pluviometr
9,2979	cintur pra cima	9,9934	sociedad educ	9,9823	recuper arroi nome
9,2979	continu la desloc	9,9934	superior	9,9823	ruid alto
9,2979	cada gesta	9,9934	vergonh pro brasil	9,9823	sonor
9,2979	diz nao	9,9934	mec	9,9823	suj meio ambient

Fonte: Elaborado pela Autora.

Tabela 17 – Atributos do Colab - Distribuição dos Scores MI

Métrica	**Todos os atributos (1089340)**	**Atributos Selecionados (99340)**
Média	0.00003853668	0.000147
Desvio Padrão	0.00023642968	0.000773
Mínimo	0.00001160660	0.000071
1º Quartil (25%)	0.00001160661	0.000072
2º Quartil (50%)	0.00003121873	0.000085
3º Quartil (75%)	0.00004371685	0.000109
Máximo	0.10252901555	0.102529

Fonte: Elaborado pela autora.

Tabela 18 – Top 20 atributos selecionados por MI no Colab

#	**Tokens**	**Score**
1	lamp	0,102529016
2	lixo	0,067290809
3	post	0,062670715
4	agua	0,055391267
5	burac	0,055022624
6	apag	0,046181188
7	esgot	0,045124152
8	estacion	0,043251338
9	vazament	0,042422989
10	queim	0,040909973
11	entulh	0,040330503
12	arvor	0,039548046
13	bueir	0,035798343
14	lamp queim	0,033214214
15	ilumin	0,032688009
16	mato	0,029021136
17	alto	0,026206226
18	carr	0,025462626
19	vazament agua	0,024374582
20	entup	0,023295543

Fonte: Elaborado pela autora.

Tabela 19 – Pesos dos Atributos Diferenciadores Colab (cont.)

	Transportes	Energia	Resíduos Sólidos	Meio Ambiente	Saúde	Água e Saneamento	Habitação	Águas Residuais	Recreação	Segurança	Economia	% acima da 2a classe
tokens \ classes	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	
lamp	5,6	9,9	6	6	6,1	6,1	6,1	6	6,1	6,1	6,1	63%
lamp queim	6,8	12	7,1	7,1	7,2	7,2	7,2	7,2	7,3	7,2	7,3	61%
lamp apag	7,2	12	7,6	7,6	7,7	7,7	7,7	7,7	7,7	7,7	7,7	60%
post lamp	8,1	13	8,5	8,4	8,6	8,6	8,6	8,5	8,6	8,6	8,6	52%
lamp apag noit	8,8	14	9,2	9,2	9,3	9,3	9,3	9,3	9,3	9,3	9,3	52%
apag noit	8,1	13	8,5	8,5	8,6	8,6	8,6	8,6	8,6	8,6	8,6	52%
lamp queim rua	9	14	9,3	9,3	9,4	9,4	9,4	9,4	9,5	9,4	9,5	50%
acend apag	9,3	14	9,7	9,7	9,8	9,8	9,8	9,8	9,8	9,8	9,8	45%
post queim	9,3	14	9,7	9,7	9,8	9,8	9,8	9,8	9,8	9,8	9,8	44%
post apag	8,4	13	8,8	8,7	8,9	8,9	8,8	8,8	8,9	8,9	8,9	44%
lamp post	8,3	13	8,7	8,6	8,8	8,8	8,7	8,7	8,8	8,8	8,8	43%
lamp aces	8,4	13	8,8	8,8	8,9	8,9	8,9	8,9	8,9	8,9	8,9	43%
estacion faix	14	9,5	9,5	9,5	9,6	9,6	9,6	9,6	9,6	9,6	9,6	43%
lamp queim frent	9,4	14	9,8	9,8	9,9	9,9	9,9	9,9	9,9	9,9	9,9	43%
agent transit	14	9,6	9,5	9,5	9,6	9,6	9,6	9,6	9,7	9,6	9,6	43%
burac grand	13	9,1	9,1	9	9,2	9,2	9,1	9,1	9,2	9,2	9,2	43%
ctb	14	9,6	9,6	9,5	9,7	9,7	9,7	9,6	9,7	9,7	9,7	42%
lamp acend	9,5	14	9,9	9,8	9,9	10	9,9	9,9	10	10	10	42%
post lamp apag	9,5	14	9,9	9,8	9,9	10	9,9	9,9	10	10	10	42%
post lamp queim	8,7	13	9,1	9,1	9,2	9,2	9,2	9,2	9,2	9,2	9,2	42%
chef fiscaliz	14	9,6	9,6	9,6	9,7	9,7	9,7	9,7	9,7	9,7	9,7	42%
aces dia	9	13	9,4	9,4	9,5	9,5	9,5	9,5	9,5	9,5	9,5	41%
sinaliz horizontal	13	9,2	9,2	9,1	9,3	9,3	9,3	9,2	9,3	9,3	9,3	41%
nittrans	12	8,4	8,4	8,4	8,5	8,5	8,5	8,5	8,5	8,5	8,5	41%
lamp queim post	9,6	14	10	9,9	10	10	10	10	10	10	10	41%
apag queim	9,6	14	10	10	10	10	10	10	10	10	10	40%
duas lamp	9,6	14	10	10	10	10	10	10	10	10	10	40%
lamp aces dia	9,6	14	10	10	10	10	10	10	10	10	10	40%
veicul estacion	11	8	8	8	8,1	8,1	8,1	8,1	8,1	8,1	8,1	39%
aces	7,6	11	7,9	7,9	8	8	8	8	8,1	8,1	8,1	39%
prefix	13	9,3	9,3	9,3	9,4	9,4	9,4	9,4	9,4	9,4	9,4	39%
queim rua	8,7	13	9,1	9,1	9,2	9,2	9,2	9,2	9,2	9,2	9,2	39%

Fonte: Elaborado pela Autora.

Tabela 20 – Pesos de Atributos Diferenciadores do Colab (cont.)

	Transportes	Energia	Resíduos Sólidos	Meio Ambiente	Saúde	Água e Saneamento	Habitação	Águas Residuais	Recreação	Segunraça	Economia	% acima da 2a classe
tokens \ classes	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	
lamp led	9,7	14	10	10	10	10	10	10	10	10	10	39%
horizontal	13	9	9	9	9,1	9,1	9,1	9,1	9,1	9,1	9,1	39%
aces durant	9	13	9,3	9,3	9,4	9,4	9,4	9,4	9,4	9,4	9,4	39%
peron	14	9,9	9,9	9,8	10	10	10	9,9	10	10	10	38%
lamp apag queim	9,8	14	10	10	10	10	10	10	10	10	10	38%
lamp branc	9,8	14	10	10	10	10	10	10	10	10	10	38%
paul peron	14	9,9	9,9	9,9	10	10	10	10	10	10	10	38%
lamp acess	9,8	14	10	10	10	10	10	10	10	10	10	37%
coloc lamp	9,8	14	10	10	10	10	10	10	10	10	10	37%
lamp queim mes	9,8	14	10	10	10	10	10	10	10	10	10	37%
aces durant dia	9,1	13	9,4	9,4	9,5	9,5	9,5	9,5	9,6	9,5	9,6	37%
lamp post queim	9,9	14	10	10	10	10	10	10	10	10	10	37%
queim mes	9,4	13	9,7	9,7	9,8	9,8	9,8	9,8	9,9	9,8	9,9	37%
queim post	9,4	13	9,7	9,7	9,8	9,8	9,8	9,8	9,9	9,8	9,9	37%
led queim	9,9	14	10	10	10	10	10	10	10	10	10	36%
post luz apag	9,9	14	10	10	10	10	10	10	10	10	10	36%
queim frent	9,1	13	9,5	9,5	9,6	9,6	9,6	9,6	9,6	9,6	9,6	36%
luz apag	8,9	13	9,3	9,3	9,4	9,4	9,4	9,4	9,4	9,4	9,4	36%
burac	7,2	5,2	5,2	5,2	5,3	5,3	5,3	5,3	5,3	5,3	5,3	36%
troc lamp	8,7	12	9	9	9,1	9,1	9,1	9,1	9,1	9,1	9,1	36%
lamp led queim	10	14	10	10	10	10	10	10	10	10	10	35%
pare	12	9,1	9,1	9,1	9,2	9,2	9,2	9,2	9,2	9,2	9,2	35%
estacion sobr calc	12	9	8,9	8,9	9	9	9	9	9,1	9	9	34%
veicul estacion sobr	14	10	10	10	10	10	10	10	10	10	10	34%
mes lamp	10	14	10	10	11	11	11	11	11	11	11	34%
post luz queim	10	14	10	10	11	11	11	11	11	11	11	34%
vari carr estacion	14	10	10	10	10	10	10	10	10	10	10	34%
apag noit rua	10	14	10	10	11	11	11	11	11	11	11	34%
lamp fica	10	14	10	10	11	11	11	11	11	11	11	34%
estacion vaga	13	9,7	9,6	9,6	9,7	9,7	9,7	9,7	9,8	9,8	9,8	34%
estacion guia	14	10	10	10	10	10	10	10	10	10	10	34%
burac bem	14	10	10	10	10	10	10	10	10	10	10	34%

Fonte: Elaborado pela Autora.

Tabela 21 – Pesos de Atributos Diferenciadores - Colab (cont.)

	Transportes	Energia	Residuos Sólidos	Meio Ambiente	Saúde	Agua e Saneamento	Habitação	Aguas Residuais	Recreação	Segunraça	Economia	% acima da 2a classe
tokens \ classes	0	1	2	3	4	5	6	7	8	9	10	
lamp post apag	10	14	10	10	11	11	11	11	11	11	11	34%
lamp vapor	10	14	10	10	11	11	11	11	11	11	11	34%
vapor metal	10	14	10	10	11	11	11	11	11	11	11	34%
semafor pedestr	13	9,4	9,4	9,4	9,5	9,5	9,5	9,5	9,5	9,5	9,5	33%
lamp ilumin	10	14	11	10	11	11	11	11	11	11	11	33%
metal branc	10	14	11	10	11	11	11	11	11	11	11	33%
vapor metal branc	10	14	11	10	11	11	11	11	11	11	11	33%
estacion sobr	11	8,5	8,5	8,5	8,6	8,6	8,6	8,6	8,6	8,6	8,6	33%
veicul abandon mes	14	10	10	10	10	10	10	10	10	10	10	33%
acend	8	11	8,4	8,3	8,5	8,5	8,4	8,4	8,5	8,5	8,5	33%
estacion faix amarel	14	10	10	10	10	10	10	10	10	10	10	33%
estacion guia rebaix	14	10	10	10	10	10	10	10	10	10	10	33%
post dest rua	10	14	11	11	11	11	11	11	11	11	11	33%
queim enfrent	10	14	11	11	11	11	11	11	11	11	11	33%
dois lad via	14	10	10	10	10	10	10	10	10	10	10	32%
faci	14	10	10	10	10	10	10	10	10	10	10	32%
lamp vapor metal	10	14	11	11	11	11	11	11	11	11	11	32%
dois post	8,6	12	9	8,9	9	9	9	9	9,1	9	9,1	32%
elev mantiqueir	10	11	11	14	11	11	11	11	11	11	11	32%
engarraf	12	9,1	9,1	9,1	9,2	9,2	9,2	9,2	9,2	9,2	9,2	32%
apag durant	9,7	13	10	10	10	10	10	10	10	10	10	32%
escurida	8,2	11	8,5	8,5	8,6	8,6	8,6	8,6	8,7	8,6	8,7	32%
desvi burac	14	10	10	10	10	10	10	10	10	10	10	32%
estacion	7,3	5,4	5,4	5,4	5,5	5,5	5,5	5,4	5,5	5,5	5,5	32%
erva passarinh	10	11	11	14	11	11	11	11	11	11	11	32%
toc music	10	11	11	14	11	11	11	11	11	11	11	32%
vari burac	12	9	9	9	9,1	9,1	9,1	9,1	9,1	9,1	9,1	32%
estacion irregular	12	8,9	8,9	8,8	8,9	9	9	8,9	9	9	9	32%
ondul	13	9,9	9,8	9,8	9,9	9,9	9,9	9,9	9,9	9,9	9,9	32%
apag post	10	14	11	11	11	11	11	11	11	11	11	32%
post acend	10	14	11	11	11	11	11	11	11	11	11	32%
faix amarel	12	8,8	8,8	8,8	8,9	8,9	8,9	8,9	8,9	8,9	8,9	31%

Fonte: Elaborado pela Autora.

Tabela 22 – Pesos de Atributos Diferenciadores - Colab (conclusão)

	Transportes	Energia	Residuos Sólidos	Meio Ambiente	Saúde	Agua e Saneamento	Habitação	Aguas Residuais	Recreação	Segunraça	Economia	
tokens \ classes	0	1	2	3	4	5	6	7	8	9	10	% acima da 2a classe
frot	14	10	10	10	10	10	10	10	10	10	10	31%
apag durant noit	9,8	13	10	10	10	10	10	10	10	10	10	31%
acess durant	10	14	11	11	11	11	11	11	11	11	11	31%
dois post lamp	10	14	11	11	11	11	11	11	11	11	11	31%
fio caid	10	14	11	11	11	11	11	11	11	11	11	31%
lamp apag frent	10	14	11	11	11	11	11	11	11	11	11	31%
post luminar	10	14	11	11	11	11	11	11	11	11	11	31%
queim frent numer	10	14	11	11	11	11	11	11	11	11	11	31%
taxa ilumin	10	14	11	11	11	11	11	11	11	11	11	31%
var lamp apag	10	14	11	11	11	11	11	11	11	11	11	31%
estacion irregul	10	7,9	7,9	7,9	8	8	8	8	8	8	8	31%
acess cadeir	13	9,6	9,6	9,5	9,6	9,7	9,6	9,6	9,7	9,7	9,7	31%
luminar	8,1	11	8,4	8,4	8,5	8,5	8,5	8,5	8,5	8,5	8,5	31%
escurida total	9,8	13	10	10	10	10	10	10	10	10	10	31%
acess dia	10	14	11	11	11	11	11	11	11	11	11	31%
acess durant dia	10	14	11	11	11	11	11	11	11	11	11	31%
fiaca caid	10	14	11	11	11	11	11	11	11	11	11	31%
lamp acend apag	10	14	11	11	11	11	11	11	11	11	11	31%
lamp continu	10	14	11	11	11	11	11	11	11	11	11	31%
lixeir che	10	11	14	11	11	11	11	11	11	11	11	31%
settr	12	9,1	9,1	9	9,2	9,2	9,2	9,1	9,2	9,2	9,2	31%
lamp ilumin public	10	14	11	11	11	11	11	11	11	11	11	30%
post apag noit	10	14	11	11	11	11	11	11	11	11	11	30%
bloqu acess cadeir	14	10	10	10	11	11	11	11	11	11	11	30%
cade eptc	14	10	10	10	11	11	11	11	11	11	11	30%
agua trat	10	11	11	11	11	14	11	11	11	11	11	30%
faix	8,6	6,5	6,5	6,5	6,6	6,6	6,6	6,6	6,6	6,6	6,6	30%
isol acust	9,5	9,9	9,9	13	10	10	10	9,9	10	10	10	30%
burac nest	14	10	10	10	11	11	11	11	11	11	11	30%
degrau	14	10	10	10	11	11	11	11	11	11	11	30%
ramp cadeir	14	10	10	10	11	11	11	11	11	11	11	30%
burac mes	13	10	10	9,9	10	10	10	10	10	10	10	30%
var lamp	9,6	13	9,9	9,9	10	10	10	10	10	10	10	30%
tod lamp	9,9	13	10	10	10	10	10	10	10	10	10	30%
faix pedestr	10	7,7	7,7	7,7	7,8	7,8	7,8	7,8	7,8	7,8	7,8	30%

Fonte: Elaborado pela Autora.

Tabela 23 – Distribuição dos Pesos nas Classes - Colab

	0		1		2
count	99340.000000	count	99340.000000	count	99340.000000
mean	12.431967	mean	12.720454	mean	12.703477
std	0.840969	std	0.995581	std	0.980338
min	4.288456	min	4.234818	min	4.209667
25%	12.160683	25%	12.385959	25%	12.359462
50%	12.671509	50%	13.079106	50%	13.052609
75%	13.076974	75%	13.484572	75%	13.458074
max	13.770121	max	14.177719	max	14.151221

	3		4		5
count	99340.000000	count	99340.000000	count	99340.000000
mean	12.917593	mean	12.748092	mean	12.744308
std	1.147549	std	1.009997	std	1.007108
min	4.125007	min	4.178457	min	4.190448
25%	12.337621	25%	12.302354	25%	12.309719
50%	13.436233	50%	13.149652	50%	13.157017
75%	13.436233	75%	13.555117	75%	13.562482
max	14.129380	max	14.248264	max	14.255629

	6		7		8
count	99340.000000	count	99340.000000	count	99340.000000
mean	12.742494	mean	12.740902	mean	12.770253
std	1.007145	std	1.009146	std	1.038585
min	4.181242	min	4.201455	min	4.179279
25%	12.294925	25%	12.436993	25%	12.327559
50%	13.142223	50%	13.130140	50%	13.174857
75%	13.547688	75%	13.535605	75%	13.580322
max	14.240835	max	14.228752	max	14.273469

	9		10
count	99340.000000	count	99340.000000
mean	12.837003	mean	12.786933
std	1.113460	std	1.056621
min	4.171482	min	4.176555
25%	12.312089	25%	12.323016
50%	13.159387	50%	13.170314
75%	13.564852	75%	13.575779
max	14.257999	max	14.268926

0 - Transportes
1 - Energia
2 - Resíduos Sólidos
3 - Meio Ambiente
4 - Saúde
5 - Água e Saneamento
6 - Habitação
7 - Águas Residuais
8 - Recreação
9 - Segunraça
10 - Economia

LENGEDA:
- **count**: número de tokens
- **mean**: média
- **std**: desvio padrão
- **min**: menor peso
- **25%**: 1º quartil
- **50%**: 2º quartil
- **75%**: 3º quartil
- **max**: maior peso

Fonte: Elaborado pela Autora.

Tabela 24 – Atributos e Pesos por Classe no Colab (cont.)

Transportes		Energia		Resíduos Sólidos	
0		1		2	
13,7701	acess estacion	14,1777	aind apag	14,1512	acab recolh
13,7701	aqui burac	14,1777	atraz igrej	14,1512	acumul lixo mato
13,7701	asfalt velh	14,1777	breu rua	14,1512	aguent sujeir
13,7701	bloqu said	14,1777	caus insegur popul	14,1512	aloj drog acondicion
13,7701	burac lama	14,1777	continu apag noit	14,1512	assim jog
13,7701	carr estacion faix	14,1777	cost luz	14,1512	botan rua
13,7701	crat pist	14,1777	deix via escur	14,1512	calc limpez
13,7701	divisa	14,1777	dess fios	14,1512	cest lixo
13,7701	elev cadeir	14,1777	esquin escurida	14,1512	coloc horari colet
13,7701	entrad said garag	14,1777	ilumin cidad	14,1512	entulh criadour
13,7701	estacion idos	14,1777	iluminaca rua	14,1512	estulh
13,7701	estacion toda	14,1777	lamp consert	14,1512	limp morador
13,7701	faix inteir	14,1777	lamp queim pert	14,1512	lixeir ceu
13,7701	faz sinaliz	14,1777	luz apag post	14,1512	mosquit deng regia
13,7701	fiscaliz linh	14,1777	luz pracinh	14,1512	muit bich aranh
13,7701	frot onibus	14,1777	nenhum funcion	14,1512	nada recolh
13,7701	guard nada	14,1777	post alto	14,1512	porquic
13,7701	imposs atravess	14,1777	post deslig	14,1512	rat foco
13,7701	onibus colet	14,1777	refletor quebr	14,1512	ter lixo
13,7701	proibica estacion	14,1777	tarif ilumin	14,1512	local chei residu
Meio Ambiente		**Saúde**		**Água e Saneamento**	
3		4		5	
14,1294	ha arvor	14,2483	abandon chei mosquit	14,2556	abastec irregul
14,1294	hipocris gent ve	14,2483	academ bairu	14,2556	agua bastant
14,1294	hoje arvor	14,2483	academ sinaps post	14,2556	agua fornec
14,1294	hoje barulh	14,2483	confirm caso zika	14,2556	agua vaz tubul
14,1294	la raiz	14,2483	conserv esgot	14,2556	boca lobo vaz
14,1294	lei fumodrom ar	14,2483	onde funcion hospital	14,2556	bomb pra funcion
14,1294	lixo queim vergonh	14,2483	area saud	14,2556	calc vazament agua
14,1294	manguezal lot	14,2483	dect	14,2556	cano jorr
14,1294	mato alto nest	14,2483	entulh criador	14,2556	consert por
14,1294	negligent falt fiscaliz	14,2483	foc zica deng	14,2556	esbanj
14,1294	nenhum poda	14,2483	infecca	14,2556	feit ligaco
14,1294	nest pais vigor	14,2483	limp propriedad	14,2556	local consert
14,1294	nest ultim dias	14,2483	mosquist	14,2556	mal feit porc
14,1294	padro permit lei	14,2483	peg zika	14,2556	potavel desperdici
14,1294	palmeir quebr	14,2483	ved prolifer mosquit	14,2556	precis agua
14,1294	pareda som	14,2483	abandon piscin	14,2556	consert imediat
14,1294	parqu deve ser	14,2483	caix agua descobert	14,2556	hidrant vaz agua
14,1294	poluent frent	14,2483	carr fumac pass	14,2556	ver vazament
14,1294	poluica quer	14,2483	peg zica	14,2556	pressa rede
14,1294	queim transmit fumac	14,2483	piscin descobert	14,2556	agua trat

Fonte: Elaborado pela Autora.

Tabela 25 – Atributos e Pesos por Classe no Colab (conclusão)

Habitação		Águas Residuais		Recreação	
6		7		8	
14,2408	abandon beir via	14,2288	abert cheir ruim	14,2735	adequ impression
14,2408	abrig sim	14,2288	adn tamp quebr	14,2735	adiant diz divulg
14,2408	alvar cert	14,2288	agua cheir fort	14,2735	alcool dia noit
14,2408	are urban	14,2288	agua chuv pra	14,2735	alem restaur
14,2408	area mesm	14,2288	agua empoc rua	14,2735	alenc trista
14,2408	banc jornal ocup	14,2288	agua esgot ceu	14,2735	algum ressac comec
14,2408	barrac ocup toda	14,2288	agua esgot foss	14,2735	ali dia estadi
14,2408	bugigang	14,2288	agua retorn	14,2735	alun pratic escol
14,2408	habitacional localiz	14,2288	agua serv foco	14,2735	amig visit
14,2408	construid meio rua	14,2288	alag bastant	14,2735	apresent prefeit
14,2408	nova favel	14,2288	alag chov fort	14,2735	arqueolog
14,2408	ocup calc mes	14,2288	alag mau cheir	14,2735	arrum prac
14,2408	papel aloj	14,2288	alem doenc	14,2735	artes pode
14,2408	peco legisl	14,2288	antes esgot	14,2735	artesa calabres
14,2408	plur habitacional	14,2288	aparent lavag rua	14,2735	atra produco cinematograf
14,2408	permit ocup irregul	14,2288	arrum bueir	14,2735	atras foco
14,2408	rua con	14,2288	atend resolv	14,2735	bel vist
14,2408	vida rua	14,2288	banh esgot	14,2735	vez mont palc
14,2408	baix dess marquis	14,2288	boc lobo enta	14,2735	vist artesa
14,2408	barrac ocup	14,2288	boeir entop	14,2735	visual rotul
Segurança		Economia			
9		10			
14,2580	abandon roub nest	14,2689	abastec veicul		
14,2580	abord band	14,2689	abert horari		
14,2580	abord vagabund	14,2689	abert total clandestin		
14,2580	acontec plen luz	14,2689	abus deve		
14,2580	agent segur	14,2689	achar finaliz		
14,2580	ameac lojist mort	14,2689	acontec estabelec obrig		
14,2580	area assalt constant	14,2689	agric rua		
14,2580	arromb bairr	14,2689	agro		
14,2580	arromb loja	14,2689	gas botija		
14,2580	ilicit ness	14,2689	gasolin exat		
14,2580	ilumin segur mesm	14,2689	iptu roub		
14,2580	local trafic consum	14,2689	irregul zona		
14,2580	moto dupl	14,2689	izopor		
14,2580	pesso armad moto	14,2689	post valor prec		
14,2580	polic cade	14,2689	valor da total		
14,2580	polic imedi	14,2689	alvar abus		
14,2580	pont drog faz	14,2689	comerci pag impost		
14,2580	pont frenquent furt	14,2689	da nota fiscal		
14,2580	post policial outr	14,2689	empres loj comerci		
14,2580	raj tiro	14,2689	proteg empres		

Fonte: Elaborado pela Autora.