UM MODELO COMPUTACIONAL PARA MINERAÇÃO DE DADOS NO FACEBOOK APLICADO A INFERÊNCIA DOS INDICADORES DE CIDADES INTELIGENTES

Gabriel da Silva Almeida[1]; Fernando Pereira de Tolêdo[2]; Silvia Silva da Costa Botelho[3]

Resumo
Neste trabalho foi proposto o estudo de técnicas e ferramentas de mineração capazes de permitir inferir o grau de inteligência de um território, a partir de postagens do Facebook. Apresenta-se um modelo para pré-processamento de postagens e obtenção de indícios (através de termos representativos) do nível de inteligência de uma cidade frente a diferentes indicadores. A partir dos termos representativos, propôs-se também prover um método capaz de classificar uma postagem quanto a um aspecto. O método foi aplicado e validado em um estudo de caso envolvendo mais de 50 mil postagens do Facebook extraídas de páginas e grupos relacionados às informações da cidade do Rio Grande/RS, entre 2012 e 2017. Os termos representativos e as classificações obtidas podem ser utilizados em etapa de pós-processamento para a obtenção de um discurso coletivo sobre a percepção do cidadão em relação ao nível de inteligência da sua cidade. Os resultados desta pesquisa apresentam indícios que o método pode ser aplicado, uma vez que as métricas de avaliação foram adequadas. Bons resultados foram obtidos, sobretudo, para os aspectos com maior número de dados anotados, o que indica que esforços terão que ser realizados para a obtenção de uma maior quantidade de amostras para treinamento.
Palavras-chave: Redes sociais; cidades inteligentes; mineração de dados.

Abstract
In this work it was proposed the study of techniques and mining tools capable of inferring the degree of intelligence of a territory, from Facebook posts. It presents a model for pre-processing of posts and obtaining evidence (through representative terms) of the level of intelligence of a city against different indicators. From the representative terms, it was also proposed to provide a method capable of classifying a post for an aspect. The method was applied and validated in a case study involving more than 50,000 Facebook posts extracted from pages and groups related to the information of the city of Rio Grande / RS between 2012 and 2017. The representative terms and ratings obtained can be used in a post-processing stage to obtain a collective discourse about the citizen's perception of the intelligence level of his city. The results of this research show that the method can be applied, since the evaluation metrics were adequate. Good results were obtained, above all, for the aspects with the highest number of annotated data, which indicates that efforts will have to be made to obtain a larger quantity of samples for training.
Keywords: Social networks; smart cities; data mining.

Introdução

Vivemos em uma sociedade que cada vez mais evolui sua forma de acesso à informação e comunicação. Proporcionada pelo avanço das conexões com a Internet e da evolução dos ambientes

[1] Mestrando no Programa de Pós-Graduação em Modelagem Computacional pela Universidade Federal do Rio Grande – FURG. E-mail: gabrielportoalegrers@gmail.com.
[2] Doutorando no Programa de Pós-Graduação em Modelagem Computacional pela Universidade Federal do Rio Grande – FURG. E-mail: fernando_ecom@yahoo.com.br.
[3] Professora permanente do Programa de Pós-Graduação em Modelagem Computacional da Universidade Federal do Rio Grande – FURG. E-mail: silviacb@furg.br.

onipresentes, dispostos aos usuários baseados nas potencialidades criadas com o uso de dispositivos móveis, surge um novo desenho de sociedade que resulta da intromissão das vias virtuais de comunicação e acesso à informação.

Uma área que vem ganhando espaço, baseado no aparecimento destas novas vias virtuais, são as redes sociais. Segundo Fry (2014) estas redes vêm proporcionando aos usuários um mundo rápido, instantâneo, com troca de informações a cada instante. Todos os dias, usuários do *Twitter*[4], *Instagram*[5] e *Facebook*[6] produzem um enorme volume de dados (mídias sobre o que estão fazendo e sentindo).

Diante da enorme quantidade de informações que são produzidas e trafegam diariamente nas redes sociais, as pesquisas nestas mídias têm se intensificado nos últimos anos (ZENG, CHEN, LUSCH, LI 2010; WOODALL, ZHAO, PAYNABAR, SPARKS, WILSON, 2017). McCormick, Lee, Cesare, Shojaie, and Spiro (2017) destacam que a pesquisa relacionada à análise de dados oriundos das redes sociais pode servir a propósitos distintos, podendo facilitar conversas e interação entre comunidades *online*, bem como a extração de padrões de inteligência úteis para entender fenômenos.

Segundo Golder and Macy (2012) estas pesquisas representam uma perspectiva completamente nova sobre a coleta de dados. Enquanto pesquisas necessitam que os entrevistados relembrem de comportamentos e ou sentimentos, os dados das redes sociais proporcionam a oportunidade de observar comportamentos e interações humanas em tempo real e em grande escala. Partindo deste princípio, buscou-se salientar que abordagens nesta linha de pesquisa, podem contribuir para análise do nível de inteligência das cidades.

Segundo a União Europeia (MANVILLE et al., 2014), Cidades Inteligentes (do inglês *Smart Cities*) são sistemas de pessoas interagindo e usando energia, materiais, serviços e financiamento para catalisar o desenvolvimento econômico e a melhoria da qualidade de vida. Esses fluxos de interação são considerados inteligentes por fazerem o uso estratégico de infraestrutura e serviços, de informação e comunicação com planejamento e gestão urbana para dar resposta às necessidades sociais, educacionais e econômicas da sociedade. Algumas abordagens vêm sendo realizadas de modo a analisar o nível de inteligência das cidades (KHATOUN, ZEADALLY, 2016; SYSTEMS, 2018; AFONSO et al., 2013). Contudo, nenhuma destas considera em seu escopo os fatores associados à construção de saberes gerado pela população acerca destes indicadores.

Diante deste cenário, busca-se neste trabalho explorar, por meio de um estudo de caso, o uso de técnicas de mineração de dados aplicadas a coleções de postagens em redes sociais, de forma a

[4] https://twitter.com
[5] https://www.instagram.com
[6] https://www.facebook.com

inferir o entendimento da população a respeito de um determinado tema. Mais precisamente, propor um modelo computacional baseado no uso de técnicas de classificação associadas a termos representativos para que a partir destes possam ser extraídos mapas conceituais sobre os principais indicadores que compõem o nível de inteligência de um município. O modelo foi aplicado e validado em um conjunto de mais de 50 mil postagens nas quais foram considerados 11 eixos (aspectos sociais) associados às cidades inteligentes, sendo eles: Economia, Educação, Empreendedorismo, Energia, Governança, Meio Ambiente, Mobilidade, Saúde, Segurança, Tecnologia & Inovação e Urbanismo (SYSTEM, 2017). As postagens envolvem o município do Rio Grande no período de 2012 a 2017.

Ademais, essa investigação assume o desafio tecnológico de realizar uma mineração de textos em fontes fracamente estruturadas em língua portuguesa, considerando que os dados textuais contêm pouca ou quase nenhuma estrutura. O trabalho está organizado da seguinte forma. Na seção 1 apresenta-se o referencial teórico abordando inicialmente indicadores para cidades inteligentes, seguido de conceitos, definições e técnicas de mineração de textos. Na seção 2, é apresentada a proposta metodológica para execução do estudo. Na seção 3 são apresentados os resultados e as discussões referentes ao estudo. Por fim, na seção 4, são descritas as considerações finais e trabalhos futuros.

1 Referencial teórico

1.1 Cidades Inteligentes

De acordo com Giffinger and Gudrun (2010), Cidades Inteligentes podem ser definidas pela eficiente visão de futuro em várias vertentes: economia, pessoas, governança, mobilidade, meio ambiente e qualidade de vida. Hall et al. (2000), as definem como aquelas que monitoram e integram as condições de operações de todas as infraestruturas críticas da cidade, atuando de forma preventiva para a continuidade de suas atividades fundamentais. Já para Harrison and Donnelly (2011), são cidades que fazem uso sistemático das TICs para promover a eficiência no planejamento, execução e manutenção dos serviços e infraestruturas urbanas.

Atualmente, diversos indicadores estão sendo propostos para avaliar o grau de inteligência de uma cidade. Dentre os índices mais divulgados, destaca-se o *European Smart Cities* (Europeu), metodologia que descreve a cidade inteligente por meio de seis (6) características (economia inteligente, pessoas inteligentes, governança inteligente, mobilidade inteligente, meio ambiente inteligente, vida inteligente), 31 fatores e 74 indicadores (KHATOUN, ZEADALLY, 2016). No Brasil, a *Urban Systems*, apresenta o *Ranking Connected Smart Cities* (SYSTEM, 2017), que considera o desenvolvimento de 11 eixos ou aspectos, sendo eles: Economia, Educação, Empreendedorismo, Energia, Governança, Meio Ambiente, Mobilidade, Saúde, Segurança, Tecnologia & Inovação e Urbanismo. Além disso, são utilizados 70 indicadores, os quais foram

pensados e estudados para atender cada um dos 11 eixos da pesquisa. Em Gomyde (2017) é apresentado o projeto Brasil 2030: Cidades inteligentes e humanas, o qual considera a verificação em 4 eixos (Governança e Arquitetura, Urbanismo e Antropologia, Tecnologia e Segurança) e 205 indicadores; e os estudos do brasileiro Afonso et al. (2013) que propõe, para além dos indicadores, a constituição de um índice de maturidade.

Percebe-se que as abordagens mencionadas apresentam a avaliação das cidades em termos de informações quantitativas sobre o grau de infraestrutura e suas tecnologias. Porém estas informações pouco podem expressar a verdadeira apropriação e uso por parte dos cidadãos. Por exemplo, qual a importância em dados quantitativos de linhas de ônibus monitorados se a população está descontente com este serviço?

Este entendimento poderia trazer uma riqueza de representação atualmente não capturada pelos indicadores de cidades inteligentes disponíveis. Desta forma, neste trabalho, diferente das demais abordagens, buscou-se apresentar como estes indicadores (percepções) podem ser encontrados e representados por meio de um recurso tecnológico (redes sociais), tendo em vista que estes locais tornam-se espaços onde os indivíduos manifestam comentários relacionados a experiências do seu cotidiano. Desta forma, fez-se necessário o estudo de fatores associados à mineração de dados e descoberta de conhecimento, os quais serão detalhados na próxima seção.

1.2 Mineração de textos e descoberta de conhecimento em bases de dados não-estruturados

De acordo com Lopes (2004), a mineração de textos, também conhecida como Descoberta de Conhecimento em Textos - KDT refere-se ao processo de extração de padrões interessantes e não triviais, ou conhecimento a partir de documentos em textos não-estruturados. Salton and McGill (1986) e Wives (2002) afirmam que a mineração de textos é um processo de descoberta de conhecimento que se utiliza de ferramentas computacionais e algoritmos específicos, de modo a processar e identificar textos, frases e palavras. Wives (2002) menciona que pode ser obtida por meio da aplicação de técnicas de KDD. Entretanto, KDT não inclui somente a aplicação das técnicas tradicionais de KDD, outras técnicas podem ser utilizadas.

As principais contribuições desta área estão relacionadas à busca de informações específicas em documentos, à análise qualitativa e quantitativa de grandes volumes de textos, e à melhor compreensão de textos disponíveis em documentos. Na prática, a mineração de textos define um processo inovador que auxilia na descoberta de conhecimento a partir de documentos textuais, que pode ser utilizado em diversas áreas (MORAIS, AMBRÓSIO, 2007). Conforme demonstra o estudo de Morais and Ambrósio (2007), um processo de mineração de textos pode ser dividido nas seguintes etapas: seleção de documentos, definição do tipo de abordagem dos dados (análise semântica ou

estatística), preparação dos dados, indexação e normalização, cálculo da relevância dos termos, seleção dos termos e pós-processamento (análise de resultados).

Segundo Morais and Ambrósio (2007), as principais técnicas de seleção de termos representativos são: filtragem baseada no peso do termo, seleção baseada no peso do termo, seleção por análise de coocorrência, seleção por *Latent Semantic Indexing* - LSI e seleção por análise de linguagem natural. Nesta pesquisa optou-se pela utilização de duas técnicas de análise de coocorrência entre termos nos documentos.

1.3 Técnicas de Análise de Coocorrência entre Termos em Documentos

De acordo com Carradore Sérgio, do Nascimento da Silva, and Gonçalves (2016); Baralis, Cagliero, Mahoto, and Fiori (2013), o termo correlação significa literalmente correlacionamento, sendo possível evidenciar o grau de relacionamento entre duas variáveis. Esta correlação pode ser obtida de cálculos oriundos da estatística. Entre os modelos utilizados para determinar a correlação, encontram-se a Frequência Conjunta, Média e Variância, Teste T (Manning, Manning, & Schütze, 1999), o *Chi-square* (CHURCH, MERCER, 1993), o *Phi-squared* (CHURCH, GALE, 1991), a Informação Mútua (Church & Hanks, 1990) e a *Expected Mutual Information Measure* (CHURCH, HANKS, 1990).

Neste estudo, a partir do corpus de treinamento no *dataset* anotado, de modo a obter o conjunto de termos representativos para cada aspecto, são aplicadas técnicas de análise de coocorrência entre termos nos documentos de acordo com os estudos de Sápiras (2015). Propõe-se a utilizar as técnicas *Phi-squared* (CHURCH, GALE, 1991) e *Expected Mutual Information Measure* - EMIM (CHURCH, HANKS, 1990).

1.4 Expected Mutual Information Measure

A *Expected Mutual Information Measure* (EMIM), segundo Church and Hanks (1990), consiste em uma técnica para análise de coocorrência que compara a probabilidade de observar duas palavras, *x* e *y*, junto com a probabilidade de observá-las independentemente. Basicamente consiste em um modelo baseado em coocorrência entre termos, o qual emprega técnicas para realizar a agregação da informação.

EMIM é definida por meio da Equação 1, em que *a* é número de vezes que as palavras *x* e *y* coocorrem em um documento; *b* é o número de vezes em que *x* ocorre em um documento e *y* não ocorre; *c* é o número de vezes que *y* ocorre e *x* não ocorre em um documento e, *d* é o tamanho da coleção de documentos menos o número de documentos que não contenham *x* e/ou *y*, sendo que *x* representa uma palavra candidata e *y* representa uma palavra semente (anotadas por um especialista).

$$EMIM(x,y) = log2\,\frac{a(a+b+c+d)}{(a+b)(a+c)} \quad (1)$$

Caso o resultado de EMIM seja maior que zero, *x* é considerado termo representativo. Entretanto, de acordo com Sápiras (2015), experimentos preliminares demonstram resultados insatisfatórios utilizando apenas esta condição. Ainda de acordo com Sápiras (2015), buscando melhores resultados, pode-se mesclar o resultado de EMIM com o resultado da quantidade de vezes que *x* e *y* coocorrem em todos os documentos de determinado aspecto da base de treinamento e através de experimentos, definir um número mínimo de coocorrências entre os dois termos para a seleção de *x* como termo representativo.

1.5 Phi-squared

A técnica *Phi-squared* (ϕ^2) é uma medida estatística que favorece uma alta ocorrência de eventos (Church & Gale, 1991). *Phi-squared* é definida por meio da Equação 2, cujas variáveis têm o mesmo significado das utilizadas na Equação 1.

$$\phi^2 = \frac{(ad-bc)^2}{(a+b)(a+c)(b+d)(c+d)}, onde: 0 \leq \phi^2 \leq 1 \quad (2)$$

De maneira semelhante ao que pretende-se aplicar na abordagem EMIM, propõe-se a mesclar o resultado com Phi-squared para análise dos experimentos, considerando a coocorrência entre os termos e os valores obtidos nas equações.

2 Metodologia

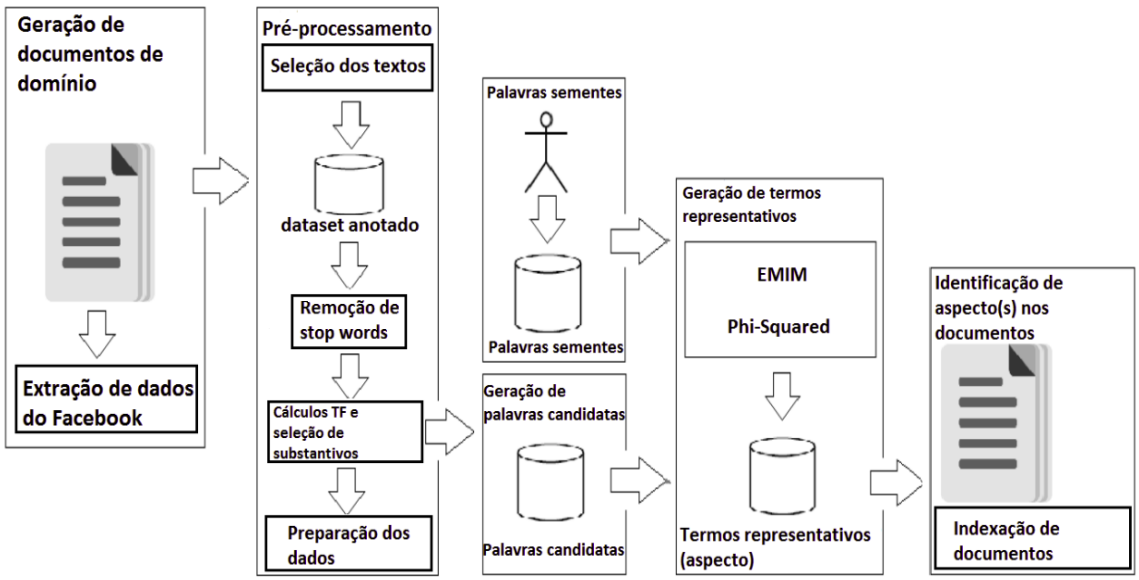
Este trabalho assume o estudo de caso como estratégia de pesquisa cuja concepção metodológica objetiva uma análise quanti-qualitativa (KING, 1994) a partir da tipologia exploratória (processo de mineração de textos) proposta por Morais e Ambrósio (2007). Buscou-se estudar técnicas e ferramentas de mineração capazes de permitir inferir o grau de inteligência de um território, a partir de postagens do Facebook. De forma mais precisa, apresenta-se um modelo e suas ferramentas capazes de pré-processar estas postagens e obter termos representativos associados aos diferentes aspectos que retratam na literatura o nível de inteligência de uma cidade.

A proposta foi aplicada em dados extraídos de 40 páginas e grupos do Facebook relacionados às informações da cidade do Rio Grande, RS, Brasil, no período que compreende a primeira mensagem de criação de cada grupo ou página até o limite de julho de 2017, sendo que o post mais antigo é datado em março de 2012. Foram obtidos um total de 53.579 mil publicações, desconsiderando todo e qualquer comentário associado.

A partir dos termos representativos, propôs-se desenvolver um método capaz de classificar uma postagem quanto a um aspecto. Os termos representativos são obtidos com base no fluxograma apresentado na Figura 1, através das técnicas EMIM e *Phi-squared*, considerando como fontes de entrada o corpus anotado de cada aspecto, além das palavras sementes e candidatas. Estes termos representativos podem ser utilizados em etapa de pós-processamento para a obtenção de um discurso

coletivo (não é objetivo deste trabalho) sobre a percepção do cidadão em relação ao nível de inteligência da sua cidade. As etapas desse processo são descritas ao longo desta seção.

Figura 1 - Fluxograma da metodologia proposta



Fonte: Elaborada pelos autores.

Na etapa de geração de documentos de domínio, extraiu-se os dados utilizando a ferramenta NodeJS Library for Facebook[7], a qual trata-se de uma implementação cliente em NodeJS para o serviço Web da Graph API Explorer, disponibilizado pelo Facebook. Os arquivos foram extraídos no formato JSON (*JavaScript Object Notation*) e migrados para a base de dados objeto-relacional MySQL, utilizando a linguagem de programação Java.

No pré-processamento, as postagens extraídas foram selecionadas. Foram considerados apenas os documentos que continham textos nas publicações com no mínimo 50 caracteres em seu conteúdo. Estes documentos selecionados constituem a entrada para as anotações através do sistema web proposto utilizando a linguagem de programação PHP[8]. Diante das anotações, gera-se um *dataset* anotado.

De posse do conjunto de treinamento, realizou-se a identificação e remoção de termos simples, tais como símbolos, caracteres especiais, datas e números, além de *stop words*, que são palavras de parada consideradas irrelevantes para o conjunto de treinamento, pois não carregam significado em seu contexto, tais como preposições e artigos.

Após a identificação e remoção das *stop words* e termos simples, procedeu-se a realização dos cálculos de frequência dos termos nos documentos. Nesta etapa, foi realizado um processo de *tokenização* nos textos, que são técnicas utilizadas para dividir esses textos em sentenças e termos.

[7] https://www.npmjs.com/package/fb.
[8] http://php.net.

Para este processo recomenda-se a utilização da ferramenta Stanford CoreNLP[9], projeto desktop, desenvolvido pela Universidade de Stanford (EUA), para o processamento de linguagem natural, composto por um conjunto de classes e bibliotecas disponíveis na linguagem de programação Java. É importante destacar que durante o processo de *tokenização*, para cada palavra encontrada nos textos do conjunto de treinamento, armazena-se a frequência da mesma no documento.

Com o intuito de gerar e obter um conjunto de palavras candidatas para cada aspecto, a partir da identificação, anotação e seleção de todos os substantivos encontrados nos respectivos corpus de treinamento, nesta etapa final de preparação dos dados, aplicou-se um processo de *Part-Of-Speech Tagger*, utilizando a ferramenta Stanford CoreNLP. Este processo realiza a identificação de propriedades gramaticais. Entretanto, os termos analisados são traduzidos para o idioma inglês, devido à ausência de suporte desta ferramenta para o idioma português. Para a tradução dos termos, foi utilizada a API (*Application Programming Interface*) Google Translator.

Após a seleção dos textos, geração do conjunto de treinamento, remoção de *stop words* e termos simples, cálculos de frequência dos termos e anotação dos substantivos nos corpus de treinamento de cada aspecto, a preparação dos dados está concluída para as próximas etapas. Procedeu-se então para a geração de palavras canditadas, a partir dos substantivos identificados, além das anotações de palavras sementes.

As palavras sementes foram anotadas para cada aspecto com base na literatura e a partir do conhecimento de domínio. Destaca-se que o conjunto de palavras candidatas para cada aspecto, foi gerado a partir do processo *Part-Of-Speech Tagger* utilizando a ferramenta Stanford CoreNLP, a qual analisou todos os termos traduzidos para o idioma inglês, provindos da etapa anterior. Em continuidade ao processo, para geração dos termos representativos para cada um dos aspectos, nesta etapa foram aplicadas duas técnicas estatísticas de análise de coocorrência de termos nos documentos: *Phi-squared* e *Expected Mutual Information Measure* (EMIM). Ambas as técnicas, utilizam como entrada o conjunto de palavras sementes, além do conjunto de palavras candidatas e os documentos da base de treinamento.

Após obter um conjunto de termos representativos para cada aspecto, a próxima etapa deste trabalho consistiu em analisar os documentos de validação. Foi selecionado para validação um percentual que varia de 20% a 30% dos documentos do *dataset* anotado. Para os aspectos Empreendedorismo e Tecnologia & Inovação, foram separados 30%, devido ao fato destes contemplarem poucas anotações. Para os demais, foram considerados 20% dos documentos anotados. Buscou-se classificar estes documentos de validação de acordo com a identificação de termos

[9] https://stanfordnlp.github.io/CoreNLP.

representativos. Para a classificação de um documento em um aspecto, considera-se a maior quantidade encontrada de termos representativos de determinado aspecto. Em caso de empate, o documento não é classificado.

Para realizar a avaliação dos resultados, referentes às técnicas de mineração de textos, de modo a obter um índice de qualidade sobre o classificador de documentos, neste trabalho foram utilizadas como métricas: precisão, revocação, f-Score e acurácia (Aggarwal & Zhai, 2012). Após esta análise, propôs-se classificar todos os documentos da base de dados que não foram anotados durante o treinamento.

3 Resultados

A partir da ferramenta web desenvolvida na etapa de pré-processamento, obteve-se um total de 11.624 anotações, sendo que 1496 destes obtiveram anotações realizadas por no mínimo 3 pessoas. A seguinte relação (quantidade) de itens anotados para cada um dos eixos de análise de indicadores foi obtida: i) Economia:157; ii) Educação: 318; iii) Empreendedorismo: 11; iv) Energia: 32; v) Governança: 143; vi) Meio Ambiente: 143; vii) Mobilidade: 91; viii) Saúde: 83; ix) Segurança: 455; x) Tecnologia e Inovação: 10 e xi) Urbanismo: 53.

Tabela 1 – Experimentos selecionados para cada aspecto

Aspecto	Experimento	Termos	DR[1]	DNR[2]	VP[3]	FP[4]	VN[5]	FN[6]	Precisão	Revocação	F-Score	Acurácia
Economia	IV	SR[7]	31	266	22	3	263	9	8.800	7.097	7.857	9.596
Educação	VIII	R[8]	63	234	53	5	229	10	9.138	8.413	8.760	9.495
Empreendedorismo	III	SR[7]	3	294	2	0	294	1	10.0	6.667	8.0	9.966
Energia	V	R[8]	6	291	5	0	291	1	10.0	8.333	9.091	9.966
Governança	VIII	R[8]	28	269	23	11	258	5	6.765	8.214	7.419	9.461
Meio Ambiente	III	SR[7]	28	269	24	1	268	4	9.600	8.571	9.057	9.832
Mobilidade	IX	SR[7]	18	279	12	2	277	6	8.571	6.667	7.500	9.731
Saúde	IV	SR[7]	16	281	11	0	281	5	10.0	6.875	8.148	9.832
Segurança	IX	R[8]	91	206	80	7	199	11	9.195	8.791	8.989	9.394
Tecnologia & Inovação	III	S[9]	3	294	2	1	293	1	6.667	6.667	6.667	9.933
Urbanismo	VII	R[8]	10	287	6	0	287	4	10.0	6.0	7.5	9.865

Fonte: Elaborada pelos autores

[1] Documentos Relevantes; [2] Documentos Não Relevantes; [3] Verdadeiros Positivos; [4] Falsos Positivos; [5] Verdadeiros Negativos; [6] Falsos Negativos; [7] Sementes e Representativas; [8] Representativas; [9] Sementes.

Para a geração dos termos representativos, foram realizados dez experimentos diferentes aplicando as técnicas EMIM e *Phi-squared*, com o intuito de analisar o resultado das métricas de precisão, revocação, f-score e acurácia, de modo a observar a qualidade do classificador. Cada experimento classificou os documentos em três abordagens diferentes: i) por termos representativos gerados pelas técnicas propostas; ii) considerando somente palavras sementes; e iii) através da mescla entre termos representativos e palavras sementes. Foi selecionado, para cada aspecto, o experimento e os termos utilizados com os melhores resultados na métrica F-score. Esta métrica foi selecionada justamente por representar a média harmônica entre a precisão e a revocação. As métricas de avaliação para os experimentos e abordagens de termos selecionados para a classificação, são apresentados na Tabela 1.

Após aplicação do classificador sob os 53579 posts, a seguinte relação (quantidade) de itens foi obtida para cada um dos eixos: i) Economia: 8462; ii) Educação: 9098; iii) Empreendedorismo: 341; iv) Energia: 705; v) Governança: 6217; vi) Meio Ambiente: 2577; vii) Mobilidade: 2754; viii) Saúde: 1710; ix) Segurança: 9726; x) Tecnologia e Inovação: 78; xi) Urbanismo: 394 e xii) Não classificados: 11517.

De modo a facilitar a visualização e entendimento do que tem sido discutido os termos representativos mais frequentes obtidos para o aspecto Educação são apresentados no formato de nuvem de palavras, ilustrados na Figura 2.

Figura 2 – Nuvem de palavras com termos representativos mais frequentes para o aspecto Educação.

Fonte: Elaborada pelos autores.

Percebe-se que o modelo foi capaz de classificar adequadamente postagens cujo quantitativo de amostras de treinamento foi significativo. Nestes casos, o universo de treinamento conseguiu agregar os principais termos relacionados à classe. Categorias com poucos dados anotados tiveram um pior desempenho. Os resultados também permitem concluir que a abordagem baseada em termos representativos é mais rica que o simples uso de palavras sementes.

O uso de tradutores e ferramentas de *tokenização* disponíveis com o uso de API também se mostrou adequado como ferramental de implementação, conseguindo tratar as peculiaridades da língua portuguesa. Com base nos termos representativos obtidos pode-se inferir elementos importantes que compõem a percepção coletiva sobre determinado indicador.

Estudos em curso aplicarão técnicas estatísticas de coocorrência para obter implicações causais entre estes elementos, capazes de fornecer discursos sobre um mais preciso entendimento da opinião coletiva destas comunidades.

4 Considerações finais

Neste trabalho buscou-se estudar e propor uma metodologia baseada em classificação de textos para ser aplicada a postagens em redes sociais, com o intuito de obter indícios sobre a percepção coletiva da população a partir de termos representativos relacionados aos indicadores de qualidade de vida no âmbito de cidades inteligentes.

A proposta foi validada em um estudo de caso associado a dados textuais extraídos de páginas e grupos temáticos do Facebook, relacionados às informações da cidade do Rio Grande / RS. Os resultados obtidos apontam indícios do funcionamento do método. Bons resultados foram obtidos, sobretudo para os aspectos com maior número de dados anotados, o que indica que esforços terão de ser realizados para a obtenção de uma maior quantidade de amostras anotadas para treinamento. Inclusive, esta expansão do *dataset* já está sendo realizada. Como trabalhos futuros pretende-se minerar as opiniões das postagens, bem como aplicar técnicas estatísticas capazes de fornecer discursos coletivos a respeito dos temas minerados.

5 Agradecimentos

Os autores agradecem à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo suporte financeiro para realização desta pesquisa.

Referências

AFONSO, R. A. et al. Br-SCMM: Modelo Brasileiro de Maturidade para Cidades Inteligentes. **Simpósio Brasileiro De Sistemas De Informação**, 2013.

AGGARWAL, C. C.; ZHAI, C. Mining text data: **Springer Science & Business Media**. 2012.

BARALIS, E. et al. GRAPHSUM: Discovering correlations among multiple terms for graph-based summarization. **Information Sciences**, v. 249, p. 96-109, 2013.

CARRADORE SÉRGIO, M.; DO NASCIMENTO DA SILVA, T.; GONÇALVES, A. L. Descoberta de conhecimento a partir de informações não estruturadas por meio de técnicas de correlação e associação. **Em Questão**, v. 22, n. 2, 2016.

CHOWDHURY, G. G. **Introduction to modern information retrieval**. Facet publishing, 2010.

CHURCH, K. Concordances for Parallel Text**. In proceedings of Seventh Annual Conference of the UW Centre for the New OED and Text Research**. 1991.

CHURCH, K. W.; MERCER, R. L. Introduction to the special issue on computational linguistics using large corpora. **Computational linguistics**, v. 19, n. 1, p. 1-24, 1993.

CHURCH, K. W.; HANKS, P. Word association norms, mutual information, and lexicography. **Computational linguistics**, v. 16, n. 1, p. 22-29, 1990.

FRY, A. **O uso (ou é abuso?) de mídia social em ciências sociais**, 2014. Disponível em: <https://www.socialsciencespace.com/2014/02/the-use-or-is-it-abuse-of-social-media-in-social-science/>. Acessado em: 15 jan. 2018, 10:00:00.

GIFFINGER, R.; GUDRUN, H. Smart cities ranking: an effective instrument for the positioning of the cities?. **ACE: architecture, city and environment**, v. 4, n. 12, p. 7-26, 2010.

GOLDER, S.; MACY, M. Social science with social media. **ASA footnotes**, v. 40, n. 1, p. 7, 2012.

GOMYDE, A. Cidades inteligentes e humanas. **Boletim de Conjuntura**, n. 2, p. 7-9, 2017.

HALL, R. E. et al. **The vision of a smart city**. Brookhaven National Lab., Upton, NY (US), 2000.

HARRISON, C.; DONNELLY, I. A. A theory of smart cities. In: **Proceedings of the 55th Annual Meeting of the ISSS-2011, Hull, UK**. 2011.

KHATOUN, R.; ZEADALLY, S. Smart cities: concepts, architectures, research opportunities. **Communications of the ACM**, v. 59, n. 8, p. 46-57, 2016.

KING, N. **The qualitative research interview**. Sage Publications, Inc, 1994.

LOPES, M. C. S. Mineração de dados textuais utilizando técnicas de clustering para o idioma português. **Rio de Janeiro: sn**, 2004.

MANNING, C. D.; MANNING, C. D.; SCHÜTZE, H. **Foundations of statistical natural language processing**. MIT press, 1999.

MANVILLE, C. et al. **Mapping smart cities in the EU**. Europan Parliament: Bressels, 2014. Disponível em: <http://www.europarl.europa.eu/RegData/etudes/etudes/join/2014/507480/IPOL-ITRE_ET(2014)507480_EN.pdf>. Acesso em: 09 mar. 2019.

MCCORMICK, T. H. et al. Using Twitter for demographic and social science research: Tools for data collection and processing. **Sociological methods & research**, v. 46, n. 3, p. 390-421, 2017.

MORAIS, E. A. M.; AMBRÓSIO, A. P. L. Mineração de textos. **Relatório Técnico–Instituto de Informática (UFG)**, 2007.

SÁPIRAS, Leonardo Augusto. Mineração de opiniões em aspectos em fontes de opiniões fracamente estruturadas. 2015.

SYSTEM, U. Ranking connected smart cities, 2018. Disponível em: <http://www.connectedsmartcities.com.br/resultados-do-ranking-connected-smart-cities/?lang=en>. Acessado em: 31 ago. 2018, 20:05:00.

SYSTEMS, U. **Urban systems**, 2018. Disponível em: <http://www.urbansystems.com.br/>. Acessado em: 05 abr. 2018, 19:00:00.

WIVES, L. K. Tecnologias de descoberta de conhecimento em textos aplicadas à inteligência competitiva. **Exame de Qualificação EQ-069, PPGC-UFRGS**, 2002.

WOODALL, W. H. et al. An overview and perspective on social network monitoring. **IISE Transactions**, v. 49, n. 3, p. 354-365, 2017.

ZENG, D. et al. Social media analytics and intelligence. **IEEE Intelligent Systems**, v. 25, n. 6, p. 13-16, 2010.