MESTRADO EM CIÊNCIA DA COMPUTAÇÃO

DISCIPLINA DE METODOLOGIA CIENTÍFICA

Expansão automatizada de Dicionários Léxicos para a Análise de Sentimentos por meio de Computação Evolucionária

Autor: Airton Bordin Junior
Orientador: Nádia Félix Felipe da Silva
Coorientador: Celso Gonçalves Camilo Junior

5 de junho de 2017

1 Apresentação

Airton Bordin Junior, bacharel em Ciência da Computação. Cursou os primeiros 3 anos do curso na Universidade Estadual do Oeste do Paraná (UNIOESTE) campus Foz do Iguaçu, e finalizou a graduação na faculdade Anglo Americano, na mesma cidade, no ano de 2011.

Possui, também, graduação em Gestão Pública, pelo Instituto Federal de Santa Catarina, cursado por meio da UaB, no campus de Foz do Iguaçu.

Após a graduação, cursou especialização em Redes de Computadores pela Universidade Federal Tecnológica do Paraná, campus de Cornélio Procópio e MBA em Gerenciamento de Projetos pelo Centro Universitário Dinâmica das Cataratas, em Foz do Iguaçu.

Atua profissionalmente na área da computação desde 2010. Trabalhou como desenvolvedor de software, testador e, por fim, como Analista de Sistemas no Parque Tecnológico Itaipu, responsável pela área de TI do projeto de Segurança de Barragens

Lecionou 2 semestres no curso Técnico em Informática Para Internet do Pronatec, atuando nas disciplinas de Sistemas Operacionais e Segurança de Sistemas, e 1 semestre no curso de Ciência da Computação em uma faculdade local, lecionando as disciplinas de Processamento de Imagens e Sistemas Inteligentes.

É aluno regular do programa de mestrado em Informática pela Universidade Federal de Goiás, na linha de pesquisa de Inteligência Computacional e bolsista CNPq.

Prentende aprofundar o trabalho na área de Mineração de Opiniões, também chamada de Análise de Sentimentos, mais precisamente na criação e expansão automatizada de Dicionários Léxicos, alinhado com trabalhos em andamento de alguns professores da Universidade. O interesse em pesquisas acerca desse assunto vem crescendo nos últimos anos, principalmente com o aumento da produção de conteúdo na WEB, e apresenta-se como um desafio interessante e atual e que pode trazer benefícios para diversas outras áreas como, por exemplo, o setor de saúde.

2 Resumo

O aumento no número de usuários de Internet nos últimos anos teve como consequência uma crescentre produção de conteúdo por seus usuários. Frequentemente, a WEB é utilizada como plataforma para debates, opiniões, avaliações, etc. Esse fato, alinhado a facilidade de obtenção dessas informações, fez com que a área de Análise de Sentimentos, também chamada de Mineração de Opiniões, tivesse um interesse crescente por parte de pesquisadores.

Uma das formas mais utilizadas no processo de Análise de Sentimentos é a utilização de Dicionários Léxicos - conjunto de palavras e suas polaridades, geralmente definidas como positiva, negativa ou neutra. Apesar de amplamente utilizada, essa abordagem possui alguns desafios a serem superados, como a identificação do domínio do texto, por exemplo - uma palavra pode ter um significado completamente diferente, dependendo do contexto em que se encontra.

O presente trabalho tem por objetivo implementar um processo de expansão Dicionário Léxico, de forma automatizada e dependente de contexto. Esses Dicionários serão utilizados como entrada por Sistemas Classificadores, que fazem a análise de conteúdo e retornam a Orientação Semântica do mesmo.

Devido à quantidade de combinações possíveis, a expansão léxica pode ser encarada como um problema de otimização. Para resolvê-lo, serão utilizados conceitos de algoritmos bioinspirados, mais precisamente Estratégias Evolutivas. Essa abordagem busca reproduzir no sistema os processos naturais de evolução, com conceitos de sobrevivência, mutação, entre outros.

Para a proposta, serão utlizados Sistemas Classificadores disponíveis, como o SentiHealth, e que possibilitem o acesso ao seu Dicionário Léxico, objeto que será modificado pelo sistema. Para os testes, serão utilizadas opiniões previamente avaliadas por especialistas humanos quanto a sua polaridade, buscando analisar a taxa de erro e a necessidade de mudança de cada solução candidata. A ideia principal é evoluir o Dicionário a cada ciclo de processamento (chamado de geração, no contexto de Estratégias Evolucionárias), buscando uma solução otimizada.

Espera-se que o trabalho forneça uma forma consistente de criação de Dicionários Léxicos para domínios específicos, um grande desafio para a área. Esses dicionários serão importantes artefatos que poderão ser utilizados em Sistemas Classificadores, tornando a análise mais apurada e com maior acurácia.

3 Introdução

A Mineração de Opiniões, também chamada de Análise de Opiniões ou Análise de Sentimentos, é uma linha de pesquisa abrangente e que vem sendo tema de diversos trabalhos nos últimos anos. Como observado em [Liu, 2010], esse crescente interesse sobre o assunto ocorre principalmente devido ao aumento no número de usuários de Internet e o consequente crescimento da produção de conteúdo independente na rede, como opiniões, avaliações, entre outros.

Essa área de estudo tem como principal desafio a Análise de Opiniões, descritas em linguagem natural, para a identificação da polaridade implícita ou explícita no texto. Essa polaridade é, na maior parte das vezes, identificada como uma escala de pontuação de sua característica positiva, negativa ou neutra.

Uma das principais técnicas para aumentar a acurácia a Análise de Sentimentos é a utilização de Dicionários de Dados. Esses dicionários contêm palavras previamente avaliadas por especialistas humanos, principalmente quanto à sua polaridade. Neste contexto, esse conjunto de palavras, juntamente com suas polaridades, é chamado de Dicionário Léxico ou Dicionário de Sentimentos.

Porém, é evidente a limitação inerente à estratégia de utilização do Dicionário Léxico - a própria lista de palavras disponíveis. Esse fato muitas vezes limita a realização de uma análise mais profunda sobre determinado contexto. Nesse sentido, um dos principais desafios na área de Mineração de Opiniões é a criação e ampliação do Dicionário Léxico de forma automatizada, tema central do presente trabalho. Grande parte desses dicionários são construídos de forma manual, fato que caracteriza uma limitação óbvia para a maior parte dos contextos e domínios, como observado em [Duwairi et al., 2015].

Consciente dessa limitação, a ideia principal do presente trabalho é a criação de um processo automatizado de expansão de Dicionário Léxico, sensível a domínios específicos, fazendo uso de técnicas de algoritmos bioinspirados da classe de Computação Evolucionária, mais precisamente Programação Evolucionária. Resumidamente essa classe de algoritmos busca reproduzir processos naturais de evolução, com conceitos de sobrevivência, mutação, entre outros. Esse processo fará a adequação das polaridades das palavras contidas no dicionário de forma a maximizar a corretude das avaliações. Para avaliar a taxa de erro de cada solução parcial gerada, serão utilizados conjuntos conhecidos de documentos previamente avaliados por especialistas humanos, de forma a compará-los com a solução gerada pelo sistema. Para a avaliação, sistemas de classificação de sentimentos disponíveis serão utilizados.

Devido à característica intrínseca do próprio problema, serão utilizadas na pesquisa bases de textos em inglês. Ao mesmo tempo, muitas técnicas utilizadas durante o trabalho também poderão ser utilizadas para resolver problemas em português, com as devidas alterações.

Espera-se que esse processo, bem como os Dicionários Léxicos por ele gerado, possam ser utilizados como entrada de processos de avaliação em diversas áreas de Mineração de Opiniões como, por exemplo, Análise de Sentimentos em redes sociais. Algumas dessas pesquisas são realizadas na própria instituição, apoiando, assim, o trabalho de outros pesquisadores. Além disso, devido ao caráter automatizado dessa solução proposta, o mesmo processo poderá ser utilizado, avaliado e melhorado para outras situações, contextos e idiomas.

Por fim, a pesquisa e a utilização de diversas técnicas de PNL e expansão automatizada de Dicionário Léxico poderão servir como um *benchmark* dos principais métodos e classificadores, auxiliando na escolha de ferramentas e abordagens para trabalhos futuros em contextos específicos.

Para apoiar a clareza e desenvolvimento da proposta, o presente documento está estruturado da seguinte forma: o capítulo 4 tratará da descrição do problema, principais limitações e dificuldades no contexto de Análise de Sentimentos. O capítulo 5 tratará dos objetivos gerais e específicos. A revisão bibliográfica, apresentada no capítulo 6, tem por objetivo criar o embasamento teórico para apoiar nas soluções propostas, apresentando o estado da arte sobre o assunto, bem como a definição de conceitos fundamentais. O capítulo 7 apresenta o impacto científico da solução e suas possíveis contribuições para a área. A metodologia, descrita no capítulo 8, descreve a forma como serão desenvolvidas cada uma das etapas do processo, seguida do capítulo com uma previsão de cronograma do trabalho. Resultados esperados são descritos no capítulo 10, seguidos da identificação dos colaboradores e participantes do projeto e, por fim, as referências bibliográficas utilizadas na proposta.

4 Descrição do Problema

Como discutido na seção anterior, uma das principais técnicas utilizadas para a Análise de Sentimentos faz uso de um Dicionário Léxico de palavras e suas polaridades, geralmente classificadas como positiva, negativa ou neutra. Essa abordagem, apesar de trazer benefícios ao processo de Mineração de Opiniões, possui algumas limitações. A principal delas está justamente ligada à disponibilidade das palavras, bem como sua correta polaridade. Outra dificuldade encontrada no uso de dicionários vem do fato que, em diferentes domínios, uma palavra pode ter um significado e, até mesmo, uma força sentimental diferente. A palavra "câncer", por exemplo, em um contexto médico, pode não ter uma conotação negativa (muitas vezes é uma palavra neutra), diferente de outros contextos.

Construir Dicionários Léxicos para domínios específicos é mais complexo que a construção de conjunto de palavras independentes de contexto, como cita [Kanayama and Nasukawa, 2006].

Buscando resolver esses problemas - falta de palavras no Dicionário Léxico e criação de Dicionários para contextos específicos - muitas técnicas foram desenvolvidas nos últimos anos. A maior parte delas utiliza a estrutura sintática dos textos como forma de tentar encontrar a polaridade mais adequada a um conjunto de palavras. Dentre as principais técnicas podemos citar o Pointwise Mutual Information, apresentado por [Turney, 2002], coerência do contexto, discutido em [Kanayama and Nasukawa, 2006], entre outros.

Uma limitação às soluções anteriores vem do fato da dificuldade inerente em trabalhar com dados não estruturados, neste contexto, linguagem natural. Esse fato é agravado quando estamos trabalhando com textos em redes sociais, por possuírem um caráter informal, contendo abreviações, gírias, trocadílhos, etc. Portanto, não é trivial expandir um Dicionário Léxico, principalmente para domínios específicos, fazendo uso das estruturas sintáticas e semânticas dos textos.

Determinar a polaridade correta da palavra, para cada contexto, portanto, apresenta-se como um desafio para as pesquisas na área. Mesmo as principais técnicas utilizadas demandam uma validação manual considerável e, em alguns contextos específicos, como o político, não atingem resultados satisfatórios.

Testar todas as possibilidades para a criação de um Dicionário Léxico perfeito, ou com o menor erro possível, também torna-se inviável, pois demandaria a resolução de um problema combinatório com muitas variáveis, o que demandaria um tempo exponencial de processamento, caracterizando-se como um problema NP-completo, ou seja, não solucionável em tempo polinomial.

Uma forma eficiente de resolver essa classe de problemas é a utilização de algoritmos bioinspirados. Por sua característica inerentemente paralela, esse tipo de abordagem consegue encontrar soluções ótimas para problemas complexos com grandes espaços de busca. Apesar de, geralmente, iniciar a busca da solução de forma aleatória, processos de modificação e adequação do algoritmo (nesse contexto chamado de operadores genéticos) permitem que as melhores soluções sejam selecionadas e evoluídas, de forma a maximizar o resultado final do processo.

Para testar a acurácia dos resultados desse processo, é necessário realizar testes em classificadores de sentimentos disponíveis, bem como a utilização de conjuntos de opiniões previamente avaliados por especialistas humanos, para comparar com os resultados da solução proposta.

5 Objetivos

5.1 Objetivo geral

O presente trabalho tem por objetivo criar um sistema para a expansão de um Dicionário Léxico, contendo palavras e suas respectivas orientações semânticas (positiva, negativa, neutra) para um determinado contexto, fazendo uso de técnicas de algoritmos bioinspirados, mais precisamente Programação Evolutiva. A criação de um Léxico abrangente e específico para um determinado contexto é um desafio para a área de Análise de Sentimentos, e fundamental para o correto funcionamento de todo o processo de análise dos dados.

5.2 Objetivos especificos

1. Criar de um sistema para a expansão automatizada de dicionário léxico para domínios específicos;
2. Criar Dicionários Léxicos consolidados para domínios específicos, prontos para serem utilizados por sistemas de Análise de Sentimentos;
3. Analisar comparativamente a técnica proposta com outras técnicas da literatura, de forma a apoiar a evolução de soluções existentes;
4. Publicar trabalhos sobre o assunto de forma a expandir o conhecimento sobre a utilização de algoritmos bioinspirados na área de Análise de Sentimentos.

6 Revisão bibliográfica

O aumento considerável na quantidade de conteúdo disponível na WEB nos últimos anos tornou possível o acesso a grandes e valiosas bases de dados e informações, como afirma [Guimaraes et al., 2016]. A crescente utilização de redes sociais e o consequente aumento no número de compartilhamento de opiniões pessoais motivaram o interesse crescente na área de Análise de Sentimentos, também chamada de Mineração de Sentimentos ou Mineração de Opiniões.

A Análise de Sentimentos é uma linha de pesquisa multidisciplinar, podendo ser considerada uma subárea de Processamento de Linguagem Natural (PNL), como afirma [Liu, 2010]. O autor, um dos principais nomes sobre o assunto, conceitualiza o problema e propõe uma forma estruturada de organização dos dados não estruturados, característica instínseca dos textos em linguagem natural, objeto de entrada da pesquisa. A definição de opinião como uma quíntupla (entidade, aspecto da entidade, sentimento, autor e tempo) é utilizada em grande parte dos trabalhos na área, caracterizando-se, portanto, como elemento fundamental nas pesquisas sobre o assunto. Visão geral sobre o tema e principais desafios e técnicas são vistos também em [Mohammad, 2016], [Ghaleb and Vijendran, 2016], [Guimaraes et al., 2016], [Taboada et al., 2011], [Bandhakavi et al., 2016], [D'Andrea et al., 2015], entre outros trabalhos.

Uma das formas mais comuns para realizar a Análise de Sentimentos, conforme argumenta [Guimaraes et al., 2016], é por meio da utilização de um Dicionário Léxico (algumas vezes chamado de Dicionário de Sentimentos), um conjunto de palavras e suas orientações semânticas (também chamadas de polaridades), frequentemente representadas como positiva, negativa ou neutra. A obtenção de um Dicionário consistente é essencial para uma correta Mineração de Sentimentos.

Ainda em [Guimaraes et al., 2016], observamos que existem, basicamente, 3 formas de criação e expansão de um Dicionário Léxico: manual - processo realizado por especialistas humanos que analisam cada palavra, atribuindo uma Orientação Semântica para cada uma delas - e duas formas (semi) automatizadas: baseada em Dicionário e baseada em Corpus. Frequentemente, essas técnicas são utilizadas em conjunto, principalmente a validação manual de Dicionários criados de forma automatizada. Criações de Dicionários utilizando somente abordagem manual, por sua característica limitante, são menos utilizadas e não serão abordadas de forma mais aprofundada no decorrer deste trabalho.

No contexto de prognóstico automatizado de Orientação Semântica de palavras, um dos primeiros trabalhos apresentados foi [Hatzivassiloglou and McKeown, 1997], focando na previsão de polaridade de adjetivos.

Uma forma de prever a polaridade sentimental de palavras desconhecidas é levar em consideração aspectos sintáticos e semânticos do texto. [Turney, 2002] apresenta uma abordagem de expansão léxica fazendo uso da técnica de Pointwise Mutual Infomation (PMI), com o objetivo de calcular a co-ocorrência de palavras e, com isso, comparar a polaridade de novas palavras com outras já conhecidas. Nesse trabalho, amplamente referenciado por outras pesquisas, o autor compara o conjunto de palavras de Orientação Semântica desconhecida com as palavras *"excellent"* e *"poor"*, representando Orientações Semânticas positiva e negativa, respectivamente. Essas palavras previamente conhecidas utilizadas como base para a expansão do Dicionário são chamadas de palavras semente (*seed words*, em inglês). Como exemplo de trabalhos que utilizam o PMI para a criação e expansão do Dicionário Léxico podemos citar [Becker et al., 2013], [Zhou et al., 2014], [Pinto et al., 2007]. [Pantel and Pennacchiotti, 2006], entre outros.

Existem alguns dicionários disponíveis para se trabalhar com a Análise de Sentimentos. A maior parte das estratégias de criação e expansão de Dicionários Léxicos utiliza como base de palavras-semente o banco de dados *WordNet* - disponível em https://wordnet.princeton.edu/ - que fornece outras facilidades, como sinônimos e antônimos. Importante destacar, também, que as bases utilizadas na maior parte dos trabalhos citados consideram palavras no idioma inglês. Mesmo alguns trabalhos que abordaram idiomas diferentes fizeram uso dessas bases por meio de um processo de tradução automatizada.

Entre os principais trabalhos da área de Análise de Sentimentos podemos citar [Taboada et al., 2011], que apresenta uma abordagem de Mineração de Opiniões baseada em Léxico combinada com uma verificação manual. Esse trabalho apresenta o SO-CAL (Semantic Orientation Calculator), que usa

lista de palavras já consolidadas para a geração de dicionários com novas entradas e suas polaridades de forma não supervisionada. Durante a descrição do trabalho, apresenta conceitos de intensificação e negação, amplamente utilizadas nas técnicas de geração de novos Dicionários. Apesar de ser feita de forma automática, o autor utilizou uma etapa de verificação humana para a validação da consistência das palavras geradas pela técnica, fazendo uso de um serviço de *Mechanical Turk* da Amazon.

Na mesma linha, [Eisenstein, 2016] e [Bandhakavi et al., 2016] apresentam outros procedimentos para apoiar a Análise de Sentimentos. O primeiro apresenta uma abordagem usando a técnica de *Naive Bayes* para a classificação dos aspectos e cita problemas de estimativas de palavras e avaliação dos léxicos criados. O segundo faz uma comparação de algumas técnicas de avaliação em 4 conjuntos de dados diferentes, apresentando uma análise quantitativa do mesmo. Abordagens e comparações semelhantes, com algumas modificações no domínio e no idioma do problema abordado, podem ser vistos em [Khoo and Johnkhan, 2017], [Asghar et al., 2014] e [Ding et al., 2008].

A maior parte dos trabalhos citados trata de todo o processo de Análise de Sentimentos. Criar e expandir Dicionários Léxicos de forma automatizada, objetivo principal do presente trabalho, é tratado de forma central em [Widdows and Dorow, 2002] e [Duwairi et al., 2015]. O primeiro utiliza uma estratégia de criação e análise de uma estrutura de grafos, por meio de uma base padronizada de palavras-semente, que contém diversas entradas previamente avaliadas em suas polaridades e, também, a descrição de seus sinônimos. Apesar de fazer uma abordagem focada em substantivos, que representam os vértices do grafo, a ideia principal pode ser utilizada em outras estratégias de geração léxica automatizada que incorporem verbos, adjetivos, entre outros.

[Duwairi et al., 2015] dá uma visão geral da criação de um dicionário de palavras, usando como base *tweets* em árabe. Importante destaque desse último foi a inclusão de *emoticons* na análise, característica amplamente utilizada, principalmente, em escritas informais na Internet. O autor fez uso da técnica de PMI, citada anteriormente e originalmente apresentada em [Turney, 2002], com algumas adequações, principalmente nas entradas da comparação da co-ocorrência - em vez de utilizar palavras, fez uso de *emoticons* positivos e negativos.

Ainda tratando especificamente de Léxicos, [Kaji and Kitsuregawa, 2007] aborda uma estratégia de criação e expansão de dicionários analisando uma coleção de páginas HTML. Apesar de trabalhar com o idioma japonês, a técnica pode ser adequada para outros idiomas.

Apesar de fornecer resultados consistentes, a abordagem de Análise de Sentimentos baseada em Dicionário tem algumas limitações. Uma delas ocorre do fato de que é virtualmente impossível trabalhar com todas as palavras existentes. Esse fato se agrava quando estamos trabalhando com opiniões expressadas em redes sociais, com uma grande quantidade de gírias, abreviações, siglas, entre outros. Além disso, quanto maior o dicionário, maiores as chances da Orientação Semântica das palavras estarem inconsistentes. Outro problema nessa abordagem é a mudança de domínio de análise. Muitas vezes, uma palavra que possui uma polaridade em um contexto específico pode ser neutra em outro e, em alguns casos, pode ter sua Orientação Semântica completamente invertida em outros domínios. Como exemplo, podemos citar a palavra "câncer"que, em um contexto técnico, pode não ter uma Orientação Semântica negativa, diferentemente de outros domínios. [Guimaraes et al., 2016] e [Abbasi et al., 2008] citam inconsistências que podem ocorrer na mudança de domínio.

Determinar a polaridade correta da palavra, para cada contexto, portanto, apresenta-se como um desafio para as pesquisas na área. Mesmo as principais técnicas utilizadas, como pode ser visto em [Taboada et al., 2011], demandam uma validação manual considerável e, em alguns contextos específicos, como o político, não atingem resultados satisfatórios, conforme podemos observar em [Guimaraes et al., 2016].

A tentativa de criação de um Dicionário Léxico por meio de um processo determinístico, testando todas as diferentes possibilidades de polaridades para cada palavra, torna-se inviável, pois demandaria a resolução de um problema combinatório com muitas variáveis, o que exigiria um tempo exponencial de processamento, caracterizando-se como um problema NP-completo, ou seja, não solucionável em tempo polinomial. Portanto, podemos classificar a criação e a expansão do Dicionário Léxico como um problema de otimização.

Uma forma eficiente de resolver essa classe de problemas é a utilização de algoritmos bioinspirados. Por sua característica inerentemente paralela, esse tipo de abordagem consegue encontrar

resultados muito próximos da solução ótima (às vezes encontra a melhor resolução) para problemas complexos com grandes espaços de busca. Apesar de, geralmente, iniciar de forma aleatória, processos de modificação e adequação do algoritmo (nesse contexto chamado de operadores genéticos) permitem que as melhores soluções sejam selecionadas e evoluídas, de forma a maximizar o resultado final do procedimento. [Abbasi et al., 2008] discorre sobre alguns pontos importantes sobre o tema.

A Programação Evolutiva, da classe das soluções bioinspiradas, pode ser utilizada para tentar resolver o problema da criação de um Dicionário Léxico. A técnica, como apresentada em [Fogel, 2000], busca reproduzir processos naturais de evolução, usando conceitos como sobrevivência, mutação, entre outros. [Bäck et al.,] faz uma abordagem comparativa de algumas Estratégias Evolutivas, trazendo conceitos importantes e definições matemáticas. [Silveira et al., 2014] discorre sobre a utilização de Programação Evolucionária no contexto de Banco de Dados, mas os conceitos podem ser utilizados em outras áreas.

Nessa abordagem, as polaridades de cada palavra do Dicionário são definidas como indivíduos do sistema e serão avaliadas quanto à sua adequação, ou seja, a diferença entre o valor esperado e o valor real do indivíduo. A função de avaliação, responsável pelo cálculo desse teste, é chamada de função *fitness*. A cada ciclo, também chamado de geração, teremos uma solução candidata, ou seja, uma possível solução otimizada para o problema.

No presente trabalho, o cálculo do *fitness* de cada solução candidata será feito comparando um resultado conhecido, utilizando um Classificador de Sentimentos disponível para utilização, com o valor da solução candidata. Sistemas Classificadores, como o criado em [Rodrigues et al., 2016], recebem como entrada uma opinião e retornam a Orientação Semântica da mesma. Para esse processo, [Rodrigues et al., 2016] faz uso de um Dicionário próprio, com grande parte das polaridades anotadas manualmente. Outros classificadores poderão ser utilizados para o cálculo da função *fitness* da solução, como os apresentados em [Pang et al., 2002], [Zhou et al., 2014], [Silva et al., 2010], [Guimaraes et al., 2016], entre outros. O requisito principal para a utilização desses sistemas para teste no presente trabalho é de que o Classificador permita a edição e manipulação do Dicionário Léxico próprio, que será modificado pelo processo de Programação Evolucionária.

Outros trabalhos que abordam a Análise de Sentimentos, fazendo uso de Estratégias Evolutivas, podem ser vistos em [Ferreira et al., 2015], [Vohra and Teraiya, 2013] e [Haddi et al., 2013].

Conjuntos de dados previamente avaliados são amplamente utilizados para teste dos Sistemas de Classificação. Esses dados tem sua Orientação Semântica determinados por especialistas humanos, e servem como entrada para a comparação da saída dos classificadores, ou seja, são dados considerados corretos e consistentes. Principais conjuntos de dados disponíveis para utilização no processo de Análise de Sentimentos podem ser vistos em [Iqbal et al., 2015], [Taboada et al., 2011], entre outros.

Após os testes dos Dicionários Léxicos gerados por meio do processo proposto neste trabalho, serão realizadas medições e comparações estatísticas, principalmente fazendo a comparação com trabalhos anteriores em atributos como acurácia, por exemplo. Esses dados serão analisados para que seja possível avaliar possíveis ganhos com relação a outras abordagens de geração automatizada de Léxicos. Exemplos de tabulação de dados de testes e análises estatísticas, como acurácia, podem ser vistos em [Taboada et al., 2011], [Kanayama and Nasukawa, 2006], [Haddi et al., 2013].

Os Dicionários Léxicos criados pela solução proposta neste trabalho, após avaliados quanto sua acurácia, ficarão disponíveis para utilização em processos de Análise de Sentimentos nos contextos e domínios para os quais foram criados. Apesar do processo de criação do Dicionário poder levar um tempo considerável de processamento (algumas horas, dependendo da quantidade de palavras e interações), o processo será realizado somente uma vez, de modo *offline*, não impactando na utilização dos sistemas de classificação, como observa [Abbasi et al., 2008].

7 Impacto Científico

A técnica de Análise de Sentimento, por meio de um Dicionário Léxico, é uma das mais utilizadas na literatura. Mostra-se, portanto, essencial a obtenção de um conjunto de palavras consolidado, juntamente com as orientações semânticas respectivas. Uma palavra pode ter um significado e, consequentemente, uma polaridade diferente, dependendo do contexto no qual está inserido.

Um conjunto de palavras e polaridades inadequadas leva a análises inconsistentes, prejudicando o resultado final do sistema.

A solução proposta neste trabalho criará, de forma automatizada, Léxicos para diferentes domínios, que poderão servir como entrada para diversos classificadores e sistemas de análise de sentimentos. Além disso, a técnica pode ser utilizada em outros idiomas, de forma a suprir uma carência de dicionários consistentes em linguagens pouco conhecidas.

O conjunto de palavras gerado pela solução proposta poderá ser utilizado como *benchmark* para outros trabalhos na área, bem como ser expandido com outras técnicas adequadas.

Técnicas de algoritmos bioinspirados, da classe de Algoritmos Evolucionários, serão utilizadas para a resolução do problema. A intersecção dessas áreas de conhecimento foi pouco explorada até o momento na literatura e em trabalhos realizados, caracterizando, portanto, uma nova abordagem para possíveis soluções. A utilização da estratégia proposta pode incentivar a utilização de outros métodos bioinspirados, apoiando, portanto, uma alternativa às soluções mais utilizadas, baseadas em análise sintática e semântica.

8 Metodologia

Para atingir os objetivos da pesquisa, buscas na literatura serão realizadas de forma a entender o estado da arte sobre o assunto, principais soluções e abordagens utilizadas. Esses dados serão utilizados como embasamento teórico para o desenvolvimento do trabalho.

Além disso, será feita uma pesquisa das principais ferramentas, preferencialmente livres e *open source* para a utilização nos testes da solução. Descrições mais detalhadas de cada etapa podem ser encontradas nos próximos subcapítulos.

Um sistema de expansão de Dicionário Léxico será implementado e testado com algumas entradas avaliadas previamente por especialistas humanos.

Após as fases de projeto, desenvolvimento e teste da solução, dados serão coletados para a criação de indicadores sobre o sistema, bem como comparações com soluções disponíveis na literatura.

8.1 Etapas

D1. **Revisao bibliografica:** Nesta etapa do trabalho será feita uma revisão bibliográfica com vistas a identificar o estado da arte do problema que está sendo proposto. Importante registrar que esta revisão bibliográfica seguirá os moldes propostos por [Kitchenham, 2004]. Serão consultadas as bases de dados do *IEEEXplore*, *ACM Digital Library*, *Science Direct*, *Research Gate*.

D2. **Estudo dos principais classificadores de sentimentos:** Nesta etapa será feita uma pesquisa sobre os principais classificadores disponíveis, preferencialmente livres e *open source*, utilizados para a análise de sentimentos, e que permitem a manipulação de seu dicionário. O objetivo principal desse passo é selecionar ferramentas que proporcionarão dados comparativos para teste da solução proposta.

D3. **Análise dos principais léxicos disponíveis:** Busca pelos principais conjuntos de palavras, e suas respectivas polaridades, disponíveis para utilização. Esses Dicionários Léxicos servirão como base e material de testes para a solução.

D4. **Recuperação das principais bases de opiniões anotadas disponíveis:** Busca e recuperação de bases de opiniões anotadas e consistentes, representando resultados confiáveis e corretos. Essas bases, já revisadas por especialistas humanos, servirão como parâmetro de corretude da solução proposta, bem como serão utilizadas para o cálculo de erro dos resultados obtidos.

D5. **Implementação da solução:** Implementação da solução proposta, com o objetivo da expansão de um léxico, sensível a um domínio específico, que servirá como entrada para um classificador utilizado em processos de análise de sentimentos.

D6. **Teste da solução com os classificadores selecionados:** Teste dos resultados fazendo uso dos classificadores selecionados anteriormente, de forma a obter um resultado satisfatório, minimizando a taxa de erros ao comparar com resultados consolidados e previamente revisados.

D7. **Levantamento dos dados de testes e relatórios:** Levantamento dos dados da utilização da solução, fazendo uso dos classificadores selecionados, e fazendo a comparação com outros sistemas e soluções disponíveis na literatura. Essa etapa fará a classificação e organização dos resultados, de forma a facilitar a visualização, entendimento, e auxiliar na tomada de decisões sobre o projeto.

8.2 Marcos fisicos

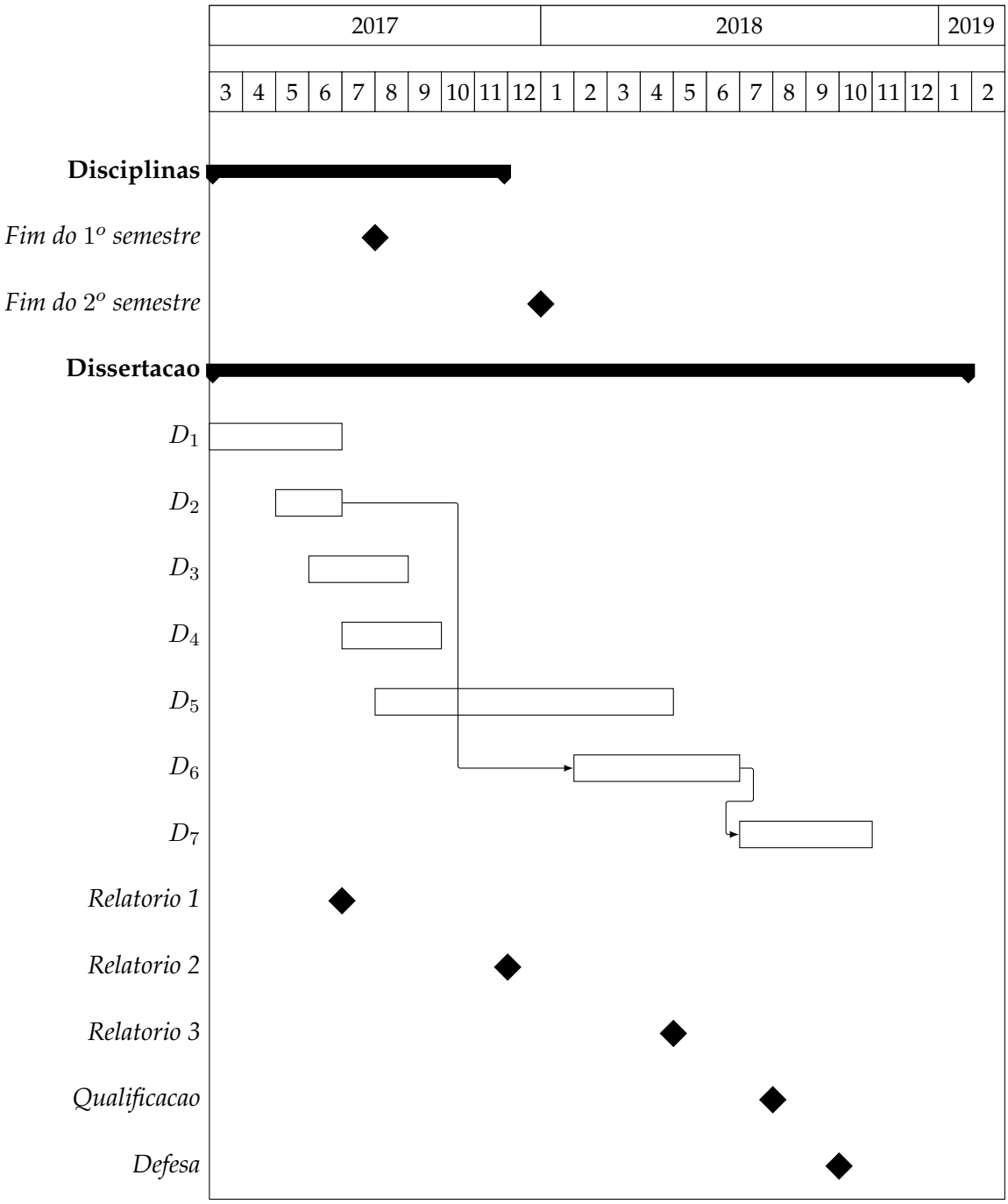
D1. Documento com a revisao bibliografica.

D2. Lista dos principais classificadores.

D3. Lista dos principais léxicos.

D4. Bases de opiniões recuperadas.

9 Cronograma de trabalho



Legenda

D1. Revisao bibliografica.

D2. Estudo dos principais classificadores de sentimentos.

D3. Análise dos principais léxicos disponíveis.

D4. Recuperação das principais bases de opiniões anotadas disponíveis.

D5. Implementação da solução.

D6. Teste da solução com os classificadores selecionados.

D7. Levantamento dos dados de testes e relatórios.

10 Resultados Esperados

Espera-se, com o presente trabalho, a criação de um processo automatizado de expansão de Dicionário Léxico dependente de domínio, fazendo uso de técnicas de algoritmos evolucionários. Nesse sentido, expansão significa tanto a criação e definição da orientação semântica de novas palavas, bem como a alteração das polaridades das palavras já existentes para um valor mais adequado ao domínio que trata o processo.

Pela característica genérica da solução, a criação de diversosLléxicos para vários domínios diferentes é limitada tão somente à escolha dos contextos específicos e à disponibilidade de dados anotados para teste da solução.

Podemos citar, também, uma possível melhoria em algumas técnicas de Análise de Sentimentos que fazem uso de Léxicos padrão, contribuindo assim para a evolução de outros sistemas de Mineração de Opiniões que usam a estratégia de dicionário.

Os resultados parciais e finais do trabalho serão descritos em artigos científicos que serão submetidos à eventos na área, de forma a compartilhar o conhecimento e avanços alcançados pela técnica proposta.

10.1 Algoritmos

Será desenvolvido um algoritmo que criará e/ou ampliará, de forma automatizada, um Dicionário Léxico para um domínio específico e que será utilizado como entrada para um sistema classificador de Análise de Sentimentos. Esse software fará uso de técnicas de algoritmos bioinspirados, mais precisamente Computação Evolucionária, para a atribuição de polaridades sentimentais para cada palavra, de forma a maximizar a taxa de acerto ao ser processado por um classificador existente. Ao passo que o algoritmo é independente de domínio, pode ser utilizado, desde que haja dados de testes suficientes, para qualquer contexto desejado.

11 Identificacao dos Participantes e Colaboradores

O presente trabalho é parte de uma pesquisa maior, realizada na UFG, que estuda a Análise de Sentimentos em todas as suas etapas.

Um projeto anterior, SentiHealth [Rodrigues et al., 2016], será utilizado como bases para testes, principalmente referente ao módulo de classificação. Além disso, o dicionário utilizado nesse sistema será manipulado pela solução proposta, servindo como entrada, portanto, do algoritmo que fará a expansão do Léxico.

Durante o trabalho, a colaboração mútua entre os participantes do grupo de pesquisa em Análise de Sentimentos será fundamental. Nesse sentido, há alunos de graduação (iniciação científica) e mestrado, que terão papel fundamental para o sucesso do tema proposto.

Por fim, mas não menos importante, a colaboração do orientador e coorientador do trabalho, ambos membros do grupo de pesquisa, será muito importante para a evolução do trabalho e para que os objetivos sejam atingidos.

12 Referencias bibliograficas

Referências

[Abbasi et al., 2008] Abbasi, A., Chen, H., and Salem, A. (2008). Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums. *ACM Trans. Inf. Syst.*, 26(3):12:1–12:34.

[Asghar et al., 2014] Asghar, M. Z., Khan, A., Ahmad, S., and Kundi, F. M. (2014). A review of feature extraction in sentiment analysis. *Journal of Basic and Applied Scientific Research*, 4(3):181–186.

[Bandhakavi et al., 2016] Bandhakavi, A., Wiratunga, N., Padmanabhan, D., and Massie, S. (2016). Lexicon based feature extraction for emotion text classification. *Pattern Recognition Letters*, pages –.

[Becker et al., 2013] Becker, L., Erhart, G., Skiba, D., and Matula, V. (2013). Avaya: Sentiment analysis on twitter with self-training and polarity lexicon expansion. In *Second Joint Conference on Lexical and Computational Semantics (* SEM)*, volume 2, pages 333–340.

[Bäck et al.,] Bäck, T., Rudolph, G., and paul Schwefel, H. Evolutionary programming and evolution strategies: Similarities and differences. In *In Proceedings of the Second Annual Conference on Evolutionary Programming*, pages 11–22.

[D'Andrea et al., 2015] D'Andrea, A., Ferri, F., Grifoni, P., and Guzzo, T. (2015). Article: Approaches, tools and applications for sentiment analysis implementation. *International Journal of Computer Applications*, 125(3):26–33. Published by Foundation of Computer Science (FCS), NY, USA.

[Ding et al., 2008] Ding, X., Liu, B., and Yu, P. S. (2008). A holistic lexicon-based approach to opinion mining. In *Proceedings of the 2008 international conference on web search and data mining*, pages 231–240. ACM.

[Duwairi et al., 2015] Duwairi, R. M., Ahmed, N. A., and Al-Rifai, S. Y. (2015). Detecting sentiment embedded in arabic social media - a lexicon-based approach. *Journal of Intelligent and Fuzzy Systems*, 29(1):107–117.

[Eisenstein, 2016] Eisenstein, J. (2016). Unsupervised learning for lexicon-based classification. *CoRR*, abs/1611.06933.

[Ferreira et al., 2015] Ferreira, L., Dosciatti, M., Nievola, J. C., and Paraiso, E. C. (2015). Using a genetic algorithm approach to study the impact of imbalanced corpora in sentiment analysis. In *FLAIRS Conference*, pages 163–168.

[Fogel, 2000] Fogel, D. B. (2000). What is evolutionary computation? *IEEE Spectr.*, 37(2):26, 28–32.

[Ghaleb and Vijendran, 2016] Ghaleb, O. A. M. and Vijendran, A. S. (2016). Survey and analysis of recent sentiment analysis schemes relating to social media. *Indian Journal of Science and Technology*, 9(41).

[Guimaraes et al., 2016] Guimaraes, N., Torgo, L., and Figueira, A. (2016). Lexicon expansion system for domain and time oriented sentiment analysis. In *Proceedings of the 8th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2016)*, pages 463–471.

[Haddi et al., 2013] Haddi, E., Liu, X., and Shi, Y. (2013). The role of text pre-processing in sentiment analysis. *Procedia Computer Science*, 17:26 – 32. First International Conference on Information Technology and Quantitative Management.

[Hatzivassiloglou and McKeown, 1997] Hatzivassiloglou, V. and McKeown, K. R. (1997). Predicting the semantic orientation of adjectives. In *Proceedings of the 35th Annual Meeting of the Association for*

Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics, ACL '98, pages 174–181, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Iqbal et al., 2015] Iqbal, M., Karim, A., and Kamiran, F. (2015). Bias-aware lexicon-based sentiment analysis. In *Proceedings of the 30th Annual ACM Symposium on Applied Computing*, SAC '15, pages 845–850, New York, NY, USA. ACM.

[Kaji and Kitsuregawa, 2007] Kaji, N. and Kitsuregawa, M. (2007). Building lexicon for sentiment analysis from massive collection of html documents. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 1075–1083, Prague, Czech Republic. Association for Computational Linguistics.

[Kanayama and Nasukawa, 2006] Kanayama, H. and Nasukawa, T. (2006). Fully automatic lexicon expansion for domain-oriented sentiment analysis. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, EMNLP '06, pages 355–363, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Khoo and Johnkhan, 2017] Khoo, C. S. and Johnkhan, S. B. (2017). Lexicon-based sentiment analysis: Comparative evaluation of six sentiment lexicons. *Journal of Information Science*, 0(0):0165551517703514.

[Kitchenham, 2004] Kitchenham, B. (2004). Procedures for performing systematic reviews. *Keele, UK, Keele University*, 33(TR/SE-0401):28.

[Liu, 2010] Liu, B. (2010). Sentiment analysis: A multifaceted problem. *IEEE Intelligent Systems*, 25(3):76–80.

[Mohammad, 2016] Mohammad, S. M. (2016). Challenges in sentiment analysis. *A Practical Guide to Sentiment Analysis, D. Das, E. Cambria, and S. Bandyopadhyay, Eds. Springer*.

[Pang et al., 2002] Pang, B., Lee, L., and Vaithyanathan, S. (2002). Thumbs up?: Sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10*, EMNLP '02, pages 79–86, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Pantel and Pennacchiotti, 2006] Pantel, P. and Pennacchiotti, M. (2006). Espresso: Leveraging generic patterns for automatically harvesting semantic relations. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, ACL-44, pages 113–120, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Pinto et al., 2007] Pinto, D., Rosso, P., and Jiménez-Salazar, H. (2007). Upv-si: Word sense induction using self term expansion. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, SemEval '07, pages 430–433, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Rodrigues et al., 2016] Rodrigues, R. G., das Dores, R. M., Camilo-Junior, C. G., and Rosa, T. C. (2016). Sentihealth-cancer: A sentiment analysis tool to help detecting mood of patients in online social networks. *International Journal of Medical Informatics*, 85(1):80 – 95.

[Silva et al., 2010] Silva, M. J., Carvalho, P., Costa, C., and Sarmento, L. (2010). Automatic expansion of a social judgment lexicon for sentiment analysis.

[Silveira et al., 2014] Silveira, B. B., Leitão-Júnior, P. S., Ramada, M. S., and Martins, B. P. (2014). Geração de base de dados para o teste de aplicações de banco de dados pelo emprego da computação evolucionária.

[Taboada et al., 2011] Taboada, M., Brooke, J., Tofiloski, M., Voll, K., and Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Comput. Linguist.*, 37(2):267–307.

[Turney, 2002] Turney, P. D. (2002). Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 417–424, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Vohra and Teraiya, 2013] Vohra, S. and Teraiya, J. (2013). A comparative study of sentiment analysis techniques. *Journal JIKRCE*, 2(2):313–317.

[Widdows and Dorow, 2002] Widdows, D. and Dorow, B. (2002). A graph model for unsupervised lexical acquisition. In *Proceedings of the 19th international conference on Computational linguistics-Volume 1*, pages 1–7. Association for Computational Linguistics.

[Zhou et al., 2014] Zhou, Z., Zhang, X., and Sanderson, M. (2014). *Sentiment Analysis on Twitter through Topic-Based Lexicon Expansion*, pages 98–109. Springer International Publishing, Cham.