MESTRADO EM CIÊNCIA DA COMPUTAÇÃO

DISCIPLINA DE METODOLOGIA CIENTÍFICA

Expansão automatizada de Léxicos para a Análise de Sentimentos por meio de programação evolucionária

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

2 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.

Sempre quis continuar estudando, e o mestrado era um objetivo a ser atingido. Por conta da falta de oportunidades na cidade onde morava (Foz do Iguaçu), esta meta teve que ser adiada. Hoje, tem a oportunidade e a honra de participar como aluno regular do programa de mestrado pela Universidade Federal de Goiás, na linha de pesquisa de Inteligência Computacional.

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 pesquisar esse 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 de conhecimento como, por exemplo, o setor de saúde.

2 Resumo

Deve ter aproximadamente 300 palavras. Alem disto, deve ter uma descricao breve de todo o projeto, atendendo a estas quatro areas:

- O que voce vai fazer? (o problema)
- Como sera feito (metodologia)
- Resultados esperados? (apenas os mais relevantes)
- Qual a importancia destes? (Conclusoes / recomendacoes)

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 utilizadas para 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.

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 Algoritmos Evolucionários, mais precisamente Programação Evolucionária. Resumidamente, o 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 próximo capítulo tratará da descrição do problema, apresentando as principais limitações e dificuldades no contexto de Análise de Sentimentos e Dicionários Léxicos. O capítulo 5 tratará dos objetivos gerais e específicos. A revisão bibliográfica, apresentada no capítulo 6, tem por objetivo o embasamento teórico para apoiar nas soluções propostas, apresentando o estado da arte sobre o assunto, bem como definindo conceitos fundamentais para entender e trabalhar com a Análise de Sentimentos. 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, como política, 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.

5 Objetivos

5.1 Objetivo geral

O presente trabalho tem por objetivo a criação de 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. Criação de um sistema para a expansão automatizada de dicionário léxico para domínios específicos;
2. Criação de dicionários léxicos consolidados para domínios específicos, prontos para serem utilizados por sistemas de Análise de Sentimentos;
3. Estudo comparativo da técnica proposta com outras técnicas da literatura, de forma a apoiar a evolução de soluções existentes;
4. Publicação de 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

[Liu, 2010] apresenta uma visão multifacetada sobre a mineração de opiniões. Neste trabalho, o autor 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].

[Taboada et al., 2011] aborda a análise e mineração de opiniões baseada em dicionários léxicos. 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 léxicos. 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 técnicas de análise de opiniões fazendo uso de dicionários léxicos. 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 criação automatizada de dicionários léxicos, tema central do presente trabalho, é tratada em sua forma geral 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 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.

A maior parte das estratégias de criação de dicionários léxicos utiliza como base de palavras semente o banco de dados *WordNet* - disponível em https://wordnet.princeton.edu/ - que fornece uma lista de palavras, sua polaridade e seus sinônimos. Importante destacar, também, que as bases utilizadas nos trabalhos supracitados consideram palavras no idioma inglês. Mesmo os trabalhos que utilizam estratégias em idiomas diferentes fizeram uso dessas bases por meio de um processo de tradução automatizada.

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 ao 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 sistemsas 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 vistas às possíveis soluções. A utilização da abordagem 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 utilizados. 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 análise, 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 sera feita uma revisao bibliografica com vistas a identificar o estado da arte do problema que esta sendo proposto. Importante registrar que esta revisao bibliografica seguira os moldes propostos por [Kitchenham, 2004]. Serao consultadas as bases de dados do [COLOCAR BASES AQUI] *Portal da Capes*, *IEEEXplore* e *ACM Digital Library*.

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

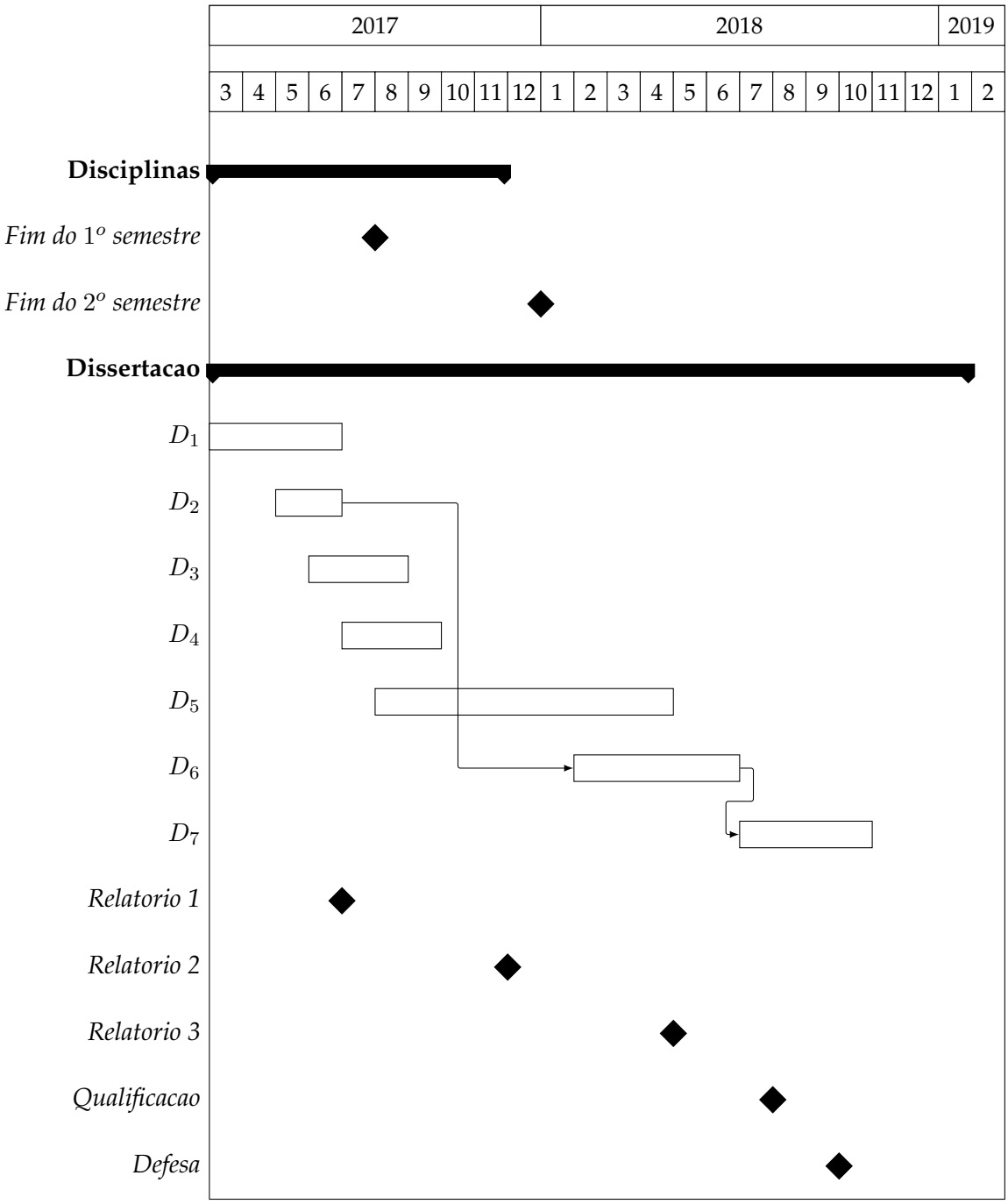
D1. Documento com a revisao bibliografica.

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 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 diversos léxicos 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 léxico para um domínio específico que será utilizado como entrada para um sistema classificador de Análise de Sentimentos. Esse software fará uso de técnicas de algoritmos bioinspirados, mais precisamente Programação Evolucionária, para a atribuição de valores 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

Aqui o candidato devera descrever se o projeto que esta sendo proposto faz parte de um projeto de pesquisa maior ou nao. Alem disto, deve descrever as possiveis colaboracoes (alunos de iniciacao cientifica, mestrado ou doutorado) que possam contribuir para o seu trabalho.

12 Referencias bibliograficas

Referências

[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 –.

[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.

[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).

[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*.

[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.

[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.