Plano de Trabalho para Iniciação Científica
PIC 2019-2020

Identificação de Linguagem Ofensiva em Mídias Sociais Utilizando Algoritmos de Aprendizado de Máquina

Murilo de Paula Araújo*

Prof. Dr. Juan Manuel Adán Coello (Orientador) **

*** Curso de Engenharia de Computação**

**** Faculdade de Engenharia de Computação**

Grupo de Pesquisa em Sistemas Inteligentes (GPqSI)

Resumo. À medida que cresce o número de pessoas que interagem através das mídias sociais multiplicam-se os casos de agressão e abuso *online,* como assédio e discursos de ódio. Tais incidentes têm trazido muitos inconvenientes aos usuários, levando alguns à desativação de contas, à autoflagelação e mesmo ao suicídio. A complexidade da questão requer uma abordagem interdisciplinar, envolvendo pesquisadores de praticamente todas as grandes áreas do conhecimento, em particular da área de computação, devido à grande quantidade e velocidade com que são geradas mensagens, *posts*, *microblogs* etc., já que é impraticável analisar a informação neles contida manualmente. Nesse contexto, o presente plano de trabalho tem o propósito de contribuir com a identificação semiautomática da presença de linguagem ofensiva em textos publicados em mídias sociais utilizando métodos de Processamento de Língua Natural (PLN) e algoritmos de aprendizado de máquina supervisionado.

Palavras-chave: linguagem ofensiva, discurso de ódio, detecção de posicionamento, mineração de opiniões, aprendizado de máquina, aprendizado profundo, *deep learning,* mídias sociais.

1 Definição do Problema

Estima-se que mais de 2,5 quintilhões de bytes de dados são criados todos os dias e que em 2020 1,7MB serão criados a cada segundo por cada pessoa na Terra[1]. Uma parte significativa desses dados corresponde a comentários e opiniões gerados em mídias sociais.

À medida em que cresce o número de pessoas que interage através das mídias sociais tem sido relatado um aumento expressivo de casos de agressão *online,* incluindo assédio de diversas naturezas e a presença de discursos de ódio. O alcance e a extensão da Internet deram a esses incidentes um poder e uma influência sem precedentes para afetar a vida de bilhões de pessoas. Casos de agressão e abuso *online* têm criado problemas de diversas naturezas e gravidade aos usuários, levado muitas deles à desativação de contas, autoflagelação e até ao suicídio. Portanto, incidentes de comportamento agressivo *on-line* tornaram-se uma importante fonte de conflito social, com potencial de produzir atividades criminosas [1]. É, portanto, oportuno que pesquisadores, administradores de redes e mídias sociais e governos, criem medidas preventivas para salvaguardar os interesses dos usuários da *web* e contribuir para a manutenção da civilidade do espaço *online*.

Uma das estratégias mais promissoras para lidar com o problema é usar métodos computacionais para identificar ofensas, agressões e incitação ao ódio em postagens, comentários, *microblogs* e demais mecanismos de comunicação *online.* Publicações e reuniões acadêmicas recentes indicam que as comunidades de pesquisadores em Processamento de Língua Natural (PLN) e Inteligência Artificial mostram interesse crescente pela questão [2][3][4][5].

O tratamento de conteúdo ofensivo pode ser dividido em três subtarefas[2]:

A. Identificação de linguagem ofensiva;
B. Categorização automática de tipos de ofensas;
C. Identificação do alvo da ofensa.

Tais subtarefas podem ser entendidas como problemas de classificação: um texto pode ser classificado como ofensivo ou não ofensivo. Por sua vez, um texto ofensivo pode conter uma ofensa explícita ou implícita, entre outros tipos, com vários níveis de gravidade, dirigido

[1] https://www.domo.com/solution/data-never-sleeps-6
[2] International Workshop on Semantic Evaluation 2019. http://alt.qcri.org/semeval2019/

a um indivíduo ou a uma comunidade (mulheres, negros, estrangeiros etc.).

O problema é desafiador, com várias e complexas perspectivas. O foco deste trabalho será a subtarefa A, partindo de algoritmos de classificação empregados em trabalhos anteriores do GPqSI que trataram de problemas relacionados (análise de sentimento e detecção posicionamento) [6][7][8][9].

2 Objetivo

O objetivo do presente plano de trabalho é produzir modelos de classificação para identificar linguagem ofensiva em textos publicados em mídias sociais, utilizando algoritmos de aprendizado de máquina supervisionado.

3 Metodologia

A busca do objetivo proposto envolverá as seguintes etapas:

1. Estudo sobre detecção de posicionamento e identificação de linguagem ofensiva em mídias sociais.
2. Estudo e aplicação de algoritmos de aprendizado de máquina.
3. Obtenção de conjuntos de dados rotulados contendo linguagem ofensiva e não ofensiva originária de mídias sociais.
4. Geração e avaliação de classificadores para a detecção de linguagem ofensiva em mídias sociais.
5. Elaboração de relatórios.

A etapa 1 consistirá da busca e estudo de artigos que discutam conceitos e soluções para os problemas de detecção de posicionamento e identificação de linguagem ofensiva em mídias sociais.

A etapa 2 envolverá o estudo de algoritmos e ferramentas no estado da arte para a geração de modelos classificadores de texto. Como parte do estudo, será reproduzido o processo empregado em trabalhos anteriores para a construção de classificadores para análise de sentimento e detecção de posicionamento. Ao término desta etapa, o bolsista deverá dominar o uso de bibliotecas e ferramentas para a construção de modelos de

classificação, como a scikit-learn[3] e a Keras[4].

Na etapa 3, serão selecionados (e eventualmente construídos) conjuntos de dados para o treinamento e avaliação dos classificadores. A construção e difusão deste tipo de conjunto de dados é uma atividade recente, voltada especialmente para a língua inglesa. Há um número restrito de iniciativas para a língua portuguesa que serão analisadas, em particular [10][11] e [12].

Na etapa 4, serão construídos modelos de classificação para a detecção de texto ofensivo, usando as ferramentas estudadas na etapa 2 e os conjuntos de dados obtidos na etapa 3. A avaliação dos modelos produzidos será feita usando as métricas *precisão*, *cobertura* e *medida-F*, assim como os tempos necessários ao treinamento e aplicação dos classificadores.

O bolsista se reunirá periodicamente com o orientador para discutir o andamento do trabalho; e registrará as atividades realizadas e resultados alcançados ao longo do desenvolvimento do plano de trabalho a fim de produzir um relatório parcial, ao final do primeiro semestre de trabalho, e um relatório final, no último mês, com o formato de um artigo científico. O bolsista participará do encontro de iniciação científica da PUC-Campinas de 2019, quando apresentará o plano de trabalho através de um pôster, e em 2020, após a conclusão do plano, quando apresentará oralmente o trabalho desenvolvido e os resultados alcançados.

4 Resultados esperados

Espera-se, ao final do desenvolvimento do plano de trabalho os seguintes resultados:

- Classificadores para a identificação de linguagem ofensiva em mídias sociais, assim como uma análise de sua acurácia.
- Conjuntos de dados para o desenvolvimento de novas pesquisas sobre o tema.
- Relatórios e artigos científicos registrando os principais aspectos e resultados do plano.

Espera-se ainda, e principalmente, que o bolsista se aproprie de métodos utilizados em pesquisas em Inteligência Artificial, particularmente na área de Aprendizado de Máquina.

[3] https://scikit-learn.org/stable/

[4] Keras: Deep Learning library for Theano and TensorFlow. https://keras.io/

5 Cronograma

Etapa	Período											
	2019					2020						
	8	9	10	11	12	1	2	3	4	5	6	7
1. Estudo sobre detecção de posicionamento e identificação de linguagem ofensiva em mídias sociais	X	X										
2. Estudo e aplicação de algoritmos de aprendizado de máquina		X	X	X								
3. Obtenção de conjuntos de dados para avaliação dos classificadores.				X	X							
4. Implementação e avaliação de classificadores.					X	X	X	X	X	X	X	
5. Elaboração de relatórios e artigos	X	X	X	X	X	X	X	X	X	X	X	X

Referências

[1] R. Kumar, G. Bhanodai, R. Pamula, M. R. Chennuru, "TRAC-1 Shared Task on Aggression Identification: IIT (ISM) COLING'18", in *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying* (TRAC-2018), 2018, p. 58–65.

[2] Z. Waseem, T. Davidson, D. Warmsley, I. Weber, "Understanding abuse: A typology of abusive language detection subtasks", *arXiv preprint* arXiv:1705.09899, 2017.

[3] D. Davidson, M. Warmsley, Macy, I. Weber, "Automated hate speech detection and the problem of offensive language", in *Eleventh International AAAI Conference on Web and Social Media*, 2017.

[4] S. Malmasi, M. Zampieri, "Challenges in discriminating profanity from hate speech", *Journal of Experimental & Theoretical Artificial Intelligence*, vol. 30, nº 2, p. 187–202, 2018.

[5] R. Kumar, A. K. Ojha, S. Malmasi, M. Zampieri, "Benchmarking aggression identification in social media", in *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, 2018, p. 1–11.

[6] F. B. Gomes, J.M. Adán Coello, F. E. Kintschner. Text Preprocessing and Ensemble Methods on Sentiment Analysis of Brazilian Portuguese Tweets. *Lecture Notes in*

Computer Science, v. 11171, p. 167-177, 2018.

[7] V. Trevisan, J. M. Adán Coello. Redes Neurais Profundas Aplicadas à Detecção de Posicionamento em Redes Sociais. In: *XXIII Encontro de Iniciação Científica e VIII Encontro de Iniciação em Desenvolvimento Tecnológico da PUC-Campinas*, 2018, Campinas.

[8] A. D. Costa Neto, J. M. ADÁN COELLO. Redes Neurais Convolucionais Aplicadas à Análise de Sentimento. In: *XXII Encontro de Iniciação Científica da PUC-Campinas*, 2017, Campinas.

[9] P. Grandin e J. M. Adán, "Piegas: A systems for sentiment analysis of tweets in portuguese", *IEEE Latin America Transactions*, vol. 14, nº 7, p. 3467–3473, 2016.

[10] R. P. de Pelle e V. P. Moreira, "Offensive Comments in the Brazilian Web: a dataset and baseline results", in *6º Brazilian Workshop on Social Network Analysis and Mining (BraSNAM 2017)*, 2017, vol. 6.

[11] R. Pelle, C. Alcântara, e V. P. Moreira, "A Classifier Ensemble for Offensive Text Detection", in *Proceedings of the 24th Brazilian Symposium on Multimedia and the Web*, 2018, p. 237–243.

[12] P. C. T. Fortuna, "Automatic detection of hate speech in text: an overview of the topic and dataset annotation with hierarchical classes", *Dissertação de Mestrado Integrado em Engenharia Informática e Computação*, Universidade do Porto, 2017.