

EXPANSÃO DE UM FRAMEWORK DE ANÁLISE DE SENTIMENTOS PARA A LÍNGUA PORTUGUESA ATRAVÉS DO USO DE TÉCNICAS DE APRENDIZADO PROFUNDO

WESLEY COSTA SILVEIRA; LARISSA ASTROGILDO DE FREITAS

Universidade Federal de Pelotas – {wcsilveira; larissa}@inf.ufpel.edu.br

1. INTRODUÇÃO

Nos últimos anos houve o advento de tecnologias com grande capacidade de processamento e de armazenamento de dados. Houve também um crescente número de usuários nas redes sociais. A partir disso, muitas empresas e pesquisadores identificaram a necessidade de analisar o que estas pessoas pensam e expressam na internet (BENEVENUTO *et al.*, 2015). Com isso, uma área que tem se desenvolvido é a Análise de Sentimentos (AS).

Conforme (BENEVENUTO *et al.*, 2015) a principal finalidade da AS é definir técnicas automáticas capazes de extrair opiniões e sentimentos com o objetivo de gerar conhecimento para um tomador de decisão. Opiniões extraídas da internet são informações relevantes, que além de proporcionarem a compreensão de eventos, também permitem prever os mesmos.

De acordo com (LIU, 2010), a AS pode ser classificada em polaridade binária ou polaridade ternária. A polaridade binária abrange apenas os sentimentos positivos e negativos, enquanto que na ternária a classificação do sentimento pode ser positiva, negativa ou neutra. Por exemplo, a frase “O dia está bonito.” é classificada como positiva, a frase “O dia está péssimo.” é classificada como negativa e a frase “Hoje é 18 segunda-feira.” é classificada como neutra.

Além disso, a AS pode ser classificada também de acordo com sua granularidade. Segundo (SILVA *et al.*, 2012), a AS pode ser realizada em três níveis de granularidade: nível de documento, nível de sentença e nível de aspecto. Quanto maior a granularidade, mais específicas se tornam as análises (BENEVENUTO *et al.*, 2015).

Este trabalho busca expandir um *framework* de AS desenvolvido por alunos da UFPel. Esse *framework* possui duas versões. A primeira versão foi desenvolvida por (PALERMO, 2019) e foi baseada em Léxicos de Sentimento para realizar as predições, enquanto que a segunda versão foi desenvolvida por (CARDOZO; FREITAS, 2020) e foi baseada em Aprendizado de Máquina.

O *framework* atual realiza suas análises em nível de documento, com polaridade binária e é específico para a língua portuguesa. A expansão do trabalho a ser desenvolvida busca melhorar os resultados de predição baseando-se em conjuntos de dados composto por *tweets* e *reviews*, que foram utilizados nos trabalhos anteriores. Para obter esse acréscimo de predição se estudará as técnicas de Aprendizado Profundo (AP), como *Transformers*, os quais têm apresentado desempenho estado da arte nas mais diversas tarefas de Processamento de Língua Natural (PLN), conforme resultados apresentados em (SOUZA *et al.*, 2020).

2. METODOLOGIA

O *framework* recebe como entrada um conjunto de dados, e classifica cada texto desse conjunto como positivo ou negativo, de acordo com a Figura 1. Essa classificação é feita de acordo com as configurações escolhidas pelo usuário, que manipula tanto as etapas de pré-processamento a serem realizadas, quanto o conjunto de dados a ser observado durante as técnicas de análise. Os módulos preenchidos em vermelho são os objetos de estudo deste trabalho.



Figura 1. Arquitetura da Expansão do *Framework* de AS para a Língua Portuguesa usando AP. Fonte: Própria.

2.1 Corpora

Atualmente os corpora utilizados no *framework* são compostos pelo TweetSentBr (BRUM; NUNES, 2018) e pelos *Reviews* de produtos eletrônicos do site Buscapé (HARTMANN et al., 2014). Neste trabalho pretende-se expandir as análises para outros corpora, a fim de fornecer um *framework* mais completo no que tange a dados de diferentes contextos/tipos.

2.2 Pré-Processamento

No módulo de pré-processamento é composta por diversas tarefas, como correção ortográfica, remoção de *stopwords*, *stemming*, *POS tagger*, *tokenizer*, detector de frases opinativas e segmentação de frases. Esses componentes possuem como finalidade o tratamento dos dados, e são usados ou não de acordo com a configuração do usuário.

2.3 Técnicas de Análise

No módulo de técnicas de análise o *framework* realiza a AS de acordo com a técnica escolhida pelo usuário, podendo ser baseada em léxico ou baseada em aprendizado de máquina. Na técnica baseada em léxico existem diferentes léxicos no *framework*, tais como: *LIWC*, *Sentilex* e *OpLexicon*. Enquanto que na técnica baseada em aprendizado de máquina pode-se destacar os algoritmos SVM (do inglês, *Support Vector Machine*) e LSTM (do inglês, *Long Short-Term Memory*).

2.3.1 BERT

O *BERT* (do inglês, *Bidirectional Encoder Representations from Transformers*), é uma técnica de AP que se baseia em *Transformers*. É o módulo a ser implementado neste trabalho com a finalidade de se obter melhores resultados de predição quando comparado com as versões anteriores do *framework*.

2.4 Saída

A saída é o último módulo, conforme apresentado na Figura 1. Nele, pode-se ter um texto (*tweets* ou *reviews*) classificado como positivo ou negativo - sentimentos que estão representados pelos *emojis*.

3. RESULTADOS E DISCUSSÃO

Como parte da revisão da literatura para posterior implementação dos módulos destacados em vermelho na Figura 1, foi elaborada a Tabela 1. Para sua construção procurou-se por trabalhos semelhantes às versões 1 e 2 do *framework* de AS para a Língua Portuguesa a ser expandido. Nela, optou-se por demonstrar apenas a métrica com o melhor resultado dentre as diferentes técnicas abordadas.

A partir disso, pode-se destacar que, ainda que o trabalho de (CARDOZO; FREITAS, 2020) apresenta resultados expressivos em termos de acurácia, ele não foi amplamente testado (considerando diferentes conjuntos de dados), o que o torna pouco informativo. Em contraposição, o trabalho de (HARTMANN et al., 2022) consiste em uma meta-análise de 272 conjuntos de dados, trazendo mais confiabilidade em suas análises. Ele destaca que, ainda que os modelos *BERT* possuam os resultados mais promissores em relação a tarefa de análise de sentimentos no nível de documento, o aspecto da interpretabilidade do modelo vem a se tornar um desafio, pois tal modelo funciona atualmente como uma caixa preta para os desenvolvedores, tornando-o pouco usual para aplicações que necessitam de uma interpretabilidade clara das análises realizadas.

Tabela 1 – Comparação de trabalhos sobre AS no nível de documento.

Trabalho	Idioma	Técnicas	Métrica	Resultado
(PALERMO, 2019)	Português	LIU, Turney	Medida-F	63,00%
(CARDOZO; FREITAS, 2020)	Português	SVM, LSTM	Acurácia	93,00%
(BALLI et al., 2022)	Turco	SVM, RL, RF, LSTM	Precisão	86,30%
(HARTMANN et al., 2022)	Inglês	RL, NB, SVM, AP	Acurácia	85,75%

4. CONCLUSÕES

Neste trabalho foi apresentado a proposta de expansão do *Framework* para AS em nível de documento para o idioma português utilizando técnicas de AP. Também foi realizada uma breve revisão da literatura de trabalhos relacionados a fim de buscar novas percepções acerca do assunto, em que pode-se destacar a dificuldade da interpretabilidade do *BERT*. Foi identificado também a necessidade de conter uma grande quantidade de dados para a etapa de pré-treinamento desse modelo. No entanto, como esse trabalho será baseado em um modelo pré-treinado do *BERT*, semelhante ao que foi descrito em (SOUZA et al., 2020), isso não virá a se tornar um problema.

5. REFERÊNCIAS BIBLIOGRÁFICAS

BENEVENUTO, F.; RIBEIRO, F.; ARAÚJO, M. Métodos para análise de sentimentos em mídias sociais. , **Brazilian Symposium on Multimedia and the Web. (Webmedia '15)**. Manaus, Brazil. October 2015.

LIU, B. Sentiment Analysis and Subjectivity. In: INDURKHYA, N.; DAMERAU, F. J. (Ed.). **Handbook of Natural Language Processing**. United States of America: Chapman and Hall/CRC, 2010. p.627–666.

SILVA, N. R.; LIMA, D.; BARROS, F. Sapair: Um processo de análise de sentimento no nível de característica. In: **INTERNATIONAL WORKSHOP ON WEB AND TEXT INTELLIGENCE (WTI'12)**, CURITIBA, 4., 2012. Anais. . . [S.l.: s.n.], 2012. p.2.

PALERMO, F. T. T. **Framework para Análise de Sentimentos em Nível de Documento**, 2019. Monografia (Graduação em Engenharia de Computação) - Curso de Engenharia de Computação, Universidade Federal de Pelotas.

CARDOZO L. S. ; FREITAS, L. A. **Expansão de um Framework de Análise de Sentimentos em Português utilizando Técnicas de Aprendizado de Máquina**, 2020. Monografia (Graduação em Ciência da Computação) - Curso de Ciência da Computação, Universidade Federal de Pelotas.

HARTMANN, J.; HEITMANN, M.; SIEBERT, C.; SCHAMP, C. More than a Feeling: Accuracy and Application of Sentiment Analysis. **SSRN Electronic Journal**, [S.l.], 05 2022.

SOUZA, F.; NOGUEIRA, R.; LOTUFO, R. BERTimbau: Pretrained BERT models for Brazilian Portuguese. In: **BRAZILIAN CONFERENCE ON INTELLIGENT SYSTEMS, BRACIS**, RIO GRANDE DO SUL, BRAZIL, OCTOBER 20-23 (TO APPEAR), 9., 2020. Anais. . . BRACIS, 2020.

BRUM, H. B.; GRAÇAS VOLPE NUNES, M. das. Building a Sentiment Corpus of Tweets in Brazilian Portuguese. In: **ELEVENTH INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION**, LREC 2018, MIYAZAKI, JAPAN, MAY 7-12, 2018, 2018. Proceedings. . . European Language Resources Association (ELRA), 2018.

HARTMANN, N. et al. A Large Corpus of Product Reviews in Portuguese: Tackling Out-Of-Vocabulary Words. In: **EUROPEAN LANGUAGE RESOURCES ASSOCIATION (ELRA)**, 2014. Proceedings. . . Reykjavik: Iceland, 2014. n.3865, p.3865–3871.

BALLI, C.; GUZEL, M. S.; BOSTANCI, E.; MISHRA, A. Sentimental Analysis of Twitter Users from Turkish Content with Natural Language Processing. **Computational Intelligence and Neuroscience**, [S.l.], v.2022, 2022.