Um Conjunto de Dados Extraído do Twitter para Análise de Sentimentos na Língua Portuguesa

Ewerton Paulo da Silva[1], Yuri Malheiros[1], Rodolffo Teles Araujo Nunes[2], Igor Leal Antunes[2], Thaís Gaudencio do Rêgo[2]

[1]Departamento de Ciências Exatas – Universidade Federal do Paraíba (UFPB)
Rio Tinto - PB - Brasil

[2]Centro de Informática - Universidade Federal do Paraíba (UFPB)
João Pessoa - PB - Brasil

{ewerton.paulo,yuri}@dcx.ufpb.br

Abstract. *The large amount of data generated by users on social networks has attracted increasing interest in the analysis of the opinions and feelings that are being expressed. For this, one of the most widely used techniques is machine learning, which needs large datasets to work properly. However, in Portuguese, few datasets for this purpose are available, limiting the development of applications in that language. Therefore, this work aims to collect messages from Twitter and to classify their sentiments to create a dataset for sentiment analysis. Volunteers labeled 2,787 messages that are publicly available. Using the dataset, we achieved 0.4503 accuracy through machine learning, a result higher than the 0.3523 accuracy using the SenticNet lexicon.*

Resumo. *A grande quantidade de dados gerada por usuários nas redes sociais tem despertado cada vez mais o interesse na análise das opiniões e sentimentos que estão sendo expressados. Para isso, uma das técnicas mais utilizadas é a aprendizagem de máquina, que precisa de grandes conjuntos de dados para funcionar adequadamente. Entretanto, na língua portuguesa, poucos conjuntos de dados para esse fim estão disponíveis, limitando o desenvolvimento de aplicações nesse idioma. Com isso, este trabalho tem como objetivo a coleta de mensagens do Twitter e a classificação do sentimento delas para criação de um conjunto de dados para a análise de sentimentos. Voluntários rotularam 2.787 mensagens que estão disponibilizadas publicamente. Utilizando os dados coletados, conseguiu-se 0,4503 de acurácia através de aprendizagem de máquina, resultado superior aos 0,3523 de acurácia usando o lexicon SenticNet.*

1. Introdução

Ao longo dos últimos anos, as redes sociais se tornaram uma das principais plataformas de comunicação do planeta, nas quais um número muito grande de pessoas consegue se expressar compartilhando diversos tipos de informações, sejam elas fotos, vídeos, textos, etc. A grande quantidade de dados gerada pelos usuários das redes sociais é preciosa e muitas vezes traz informações que não são percebidas com facilidade. Por exemplo, é possível extrair automaticamente de mensagens textuais sobre que assunto elas se referem, que idioma estão as mensagens e também que sentimento elas carregam: felicidade,

tristeza, excitação, raiva, etc. Este último tipo de informação é tratado especificamente pela área de análise de sentimentos [Liu and Zhang 2012] [Pang and Lee 2008].

A detecção de sentimentos por meio de computadores tem ganhado muita atenção nos últimos anos, tanto nas universidades quanto nas empresas [Liu et al. 2005] [Gamon 2004]. Um dos motivos de tal interesse é justamente o aumento da quantidade de conteúdo gerado pelas pessoas na Internet, principalmente quando elas estão expressando opinião. Entre as técnicas mais utilizadas para detecção de sentimentos está a aprendizagem de máquina supervisionada. Nela, classificadores usam dados previamente rotulados com os seus sentimentos para aprender padrões e conseguir prever novas entradas. Para treinar classificadores são necessários muitos dados, assim a disponibilidade de conjuntos de dados são essenciais para a realização de pesquisas e desenvolvimento de aplicações nessa área. Entretanto, conjuntos de dados com exemplos na língua portuguesa ainda são escassos, o que limita as aplicações voltadas para esse idioma [Moraes et al. 2015].

Tendo em vista esta necessidade, este trabalho tem como objetivo a coleta e classificação de *tweets*, que são as mensagens compartilhadas no Twitter, para criação de um conjunto de dados para análise de sentimentos na língua portuguesa. Para alcançar esse resultado foi desenvolvido um coletor de mensagens utilizando a API do Twitter. Em seguida, foi desenvolvida uma aplicação web para que voluntários pudessem classificar as mensagens coletadas em relação ao seu sentimento (positivo, negativo ou neutro). No total foram classificados 2.787 *tweets* sendo 888 positivos, 881 negativos e 1.018 neutros.

O restante do artigo está estruturado da seguinte forma. Na seção 2 são descritos os trabalhos relacionados, na seção 3 é apresentado o processo de criação do conjunto de dados, na seção 4 são apresentados resultados da avaliação do conjunto de dados e uma breve discussão e na seção 5 temos conclusão.

2. Trabalhos Relacionados

Outros trabalhos na literatura procuraram alcançar resultados similares criando conjuntos de dados em português para análise de sentimentos.

No trabalho de [Brum and Nunes 2017] um conjunto de dados de 15.000 mensagens rotuladas foi criado. Para isso, foi necessária a coleta de dados utilizando a API do Twitter focado em mensagens compartilhadas durante a exibição de programas de TV. As mensagens foram classificadas entre positivas, negativas e neutras. O processo de classificação foi realizado através de uma ferramenta web de anotação utilizada por sete participantes nativos da língua portuguesa com o auxílio de um guia da língua. O conjunto possui um total de 6.648 mensagens positivas, 3.926 neutras e 4.426 negativas. Também foram realizados experimentos com três métodos de aprendizagem de máquina. Ao final, o conjunto de dados criado foi disponibilizado por meio de um repositório público.

O PELESent [Corrêa et al. 2017] foi criado com o objetivo de ser um conjunto de dados com uma grande quantidade de *tweets*, possuindo um total de 980.067 mensagens. Pelo grande custo de realizar essa anotação por humanos, *emojis* foram utilizados para classificar as mensagens, sendo 554.623 positivas e 425.444 negativas. Para avaliação, métodos de classificação de polaridades foram treinados com o conjunto de dados e os modelos resultantes foram aplicados em cinco outros conjuntos de dados anotados manualmente.

O 7x1-PT é um conjunto de dados para análise de sentimentos com *tweets* que foram enviados ao longo da partida da Alemanha com o Brasil durante a Copa do Mundo de 2014 da FIFA [Moraes et al. 2015]. Durante o jogo entre as equipes foram buscadas mensagens que continham palavras relacionadas à Copa do Mundo, por exemplo, hexa, vencedor, etc. O conjunto de dados final foi classificado por dois anotadores humanos totalizando 2.728 *tweets*, sendo 157 positivos, 1.771 neutros e 800 negativos.

No trabalho de [de Arruda et al. 2015] foi criado um conjunto de dados de notícias extraídas de 4 grandes jornais brasileiros. Para coletar os dados, durante sete dias, às 20:00 horas, um *crawler* capturava pelo menos 20 notícias sobre política de perfis do Twitter dos meios de comunicação selecionados. Em seguida, essas notícias foram divididas em parágrafos para que quatro anotadores humanos determinassem sobre que pessoa o parágrafo se referia e o sentimento do parágrafo em relação a essa pessoa. No total foram classificados 1.447 parágrafos de 113 notícias.

Outro trabalho realizado abordando notícias brasileiras foi desenvolvido por [Dosciatti et al. 2015], nele foi construído um corpus de notícias para análise de sentimentos que poderiam ter as seguintes classificações: alegria, tristeza, raiva, surpresa, repugnância e medo. Foram coletados 2.000 textos que foram classificados por seis anotadores voluntários com experiência de no mínimo 15 anos em linguística.

Em comparação a esses trabalhos descritos, o nosso principal diferencial é que não restringimos o escopo das mensagens para um contexto específico e que coletamos mensagens por um período significativamente maior que os trabalhos citados. Além disso, na classificação, cada mensagem podia receber o julgamento de até cinco pessoas para obter uma maior consistência.

3. Criação do Conjunto de Dados

A criação do conjunto de dados foi realizada em duas etapas principais. Na primeira, foi desenvolvida uma ferramenta para coletar mensagens compartilhadas no Twitter. Na segunda, uma ferramenta web para rotular os textos foi desenvolvida e disponibilizada para que voluntários classificassem as mensagens coletadas de acordo com o seu sentimento (positivo, negativo ou neutro). A seguir mostraremos mais detalhes sobre como foram realizadas essas duas etapas.

3.1. Coleta de dados

Para realizar a coleta das mensagens foi desenvolvida uma ferramenta em Python que utiliza a API do Twitter para buscar mensagens compartilhadas na rede social de acordo com palavras-chave. Como centenas de milhões de *tweets* são enviados a cada dia [Domo 2019], temos uma variedade muito grande de textos compartilhados. Para buscar *tweets* com uma maior chance de ter algum sentimento, escolheu-se usar como palavras-chave adjetivos da língua portuguesa. Os adjetivos utilizados nas buscas foram disponibilizados pelo thesaurus TeP 2.0 [Dias-Da-Silva and de Moraes 2003].

A ferramenta de coleta foi executada em um servidor entre os dias 24/09/2018 e 06/12/2018. Durante esse período, a ferramenta selecionava aleatoriamente um adjetivo do thesaurus TeP 2.0 e efetuava uma busca no Twitter, guardando as mensagens encontradas. Após salvar os *tweets* no banco de dados, a ferramenta começava a checar se 30

minutos já se passaram, para em seguida escolher um novo adjetivo e recomeçar o processo. Os dados armazenados de cada *tweet* foram o ID do *tweet* e o seu conteúdo textual. Os passos do processo de coleta são sumarizados na Figura 1.

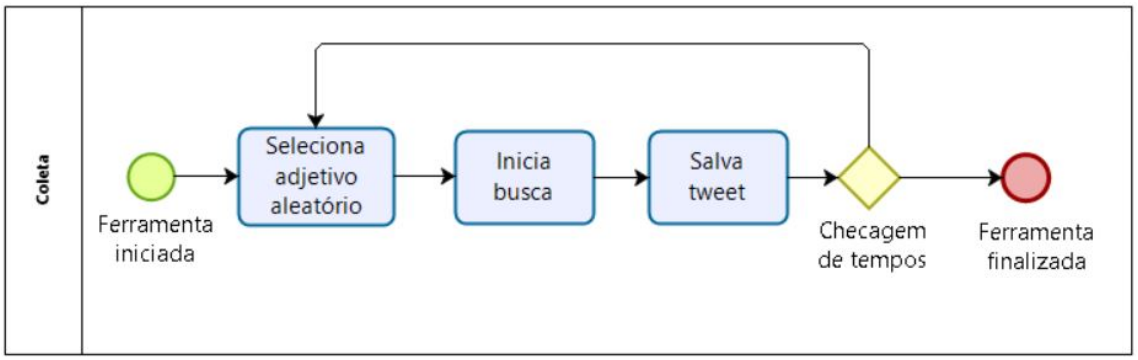

Figura 1. Passos executados para a coleta de *tweets*

A etapa de salvar *tweets* seguiu alguns critérios de avaliação do texto antes de armazená-los. Primeiramente, foi verificado se o texto do *tweet* já havia sido salvo no banco de dados. Quando isso acontecia, como não queríamos ter dados duplicados no conjunto de dados, o *tweet* era descartado. Em seguida, para tentar obter um maior número de mensagens com sentimentos foi calculada a polaridade da mensagem utilizando o SenticNet 5 [Cambria et al. 2018] através da biblioteca SenticNet API[1].

A polaridade é um número que varia no intervalo de -1 a 1. Mensagens com polaridades próximas de 0 são consideradas neutras, sem sentimento. Quanto mais próxima de -1 a polaridade de uma mensagem for, mais ela é negativa e, no caso contrário, quanto mais próxima de 1 a polaridade, mais positiva é a mensagem. Para calcular a polaridade de uma mensagem foi realizada uma média das polaridades das palavras da mensagem presentes no SenticNet. Com isso, polaridades entre -0,003 e 0,003 foram consideradas neutras no processo de coleta e, por isso, foram descartadas.

Após todo o processo de coleta, foram armazenados no banco de dados 641.471 *tweets* para serem classificados posteriormente em relação aos seus sentimentos. A Tabela 1 mostra três exemplos de *tweets* coletados.

Tabela 1. Exemplos de *tweets* coletados

Tweet
esse gel p dor é milagroso
um vampiro aventureiro e um vampiro mimado https://t.co/rVdBZaFlZv
Dividido entre a tristeza, o ódio e o veneno que existem dentro de mim

3.2. Classificação

Com os *tweets* coletados e salvos em um banco de dados, a etapa subsequente foi a classificação das mensagens por voluntários. Para isso, foi desenvolvida uma ferramenta web para que os voluntários julgassem se mensagens tinham sentimentos positivos, negativos ou neutros. Ao iniciar a ferramenta é apresentado para o usuário um *tweet* escolhido

[1]https://github.com/yurimalheiros/senticnetapi

aleatoriamente do banco de dados. Abaixo do *tweet* são disponibilizados três botões para o usuário interagir, cada um relacionado a uma das possíveis classificações. Quando o usuário seleciona um sentimento, a ferramenta grava a classificação e exibe uma nova mensagem para avaliação. A tela exibida para o usuário pode ser visualizada na Figura 2.

Figura 2. Tela da ferramenta para classificação de *tweets*

A ferramenta foi disponibilizada para os alunos dos cursos de Bacharelado em Sistemas de Informação e Licenciatura em Ciência da Computação da Universidade Federal da Paraíba classificarem as mensagens do conjunto de dados. Para tornar o processo mais confiável, um *tweet* poderia ser classificado até cinco vezes, sendo sua classificação final o sentimento mais escolhido. Assim, tentou-se dar maior consistência às classificações atribuídas, pois um julgamento destoante de um voluntário poderia ser corrigido pelos julgamentos restantes. Outro ponto importante é que foram coletados 641.471 *tweets*, um número muito grande para ser classificado pelos voluntários, portanto na ferramenta foi usado um subconjunto de 10.000 *tweets* escolhidos aleatoriamente a partir dos *tweets* coletados. Por fim, verificações manuais periódicas foram realizadas nas classificações para detecção de comportamentos indesejáveis, por exemplo, muitas atribuições de um único sentimento para muitas mensagens em pouco tempo.

Ao final da etapa de classificação foram realizadas 1.545 avaliações positivas, 1.473 negativas e 1.617 neutras. Com isso, no total foram classificados 2.787 *tweets* sendo 888 positivos, 881 negativos e 1.018 neutros. O conjunto de dados está disponível publicamente através do *GitHub* em um arquivo *CSV* que contém o ID de cada *tweet* e o seu sentimento correspondente[2]. A Tabela 2 traz uma amostra de três *tweets* que estão no conjunto de dados finais e os seus sentimentos.

[2]https://github.com/arialab/tash-pt

Tabela 2. Exemplos retirados do conjunto de dados

Texto	Sentimento
queria deixar registrado o quanto eu fiquei feliz por ter recebido um áudio da sophi falando que tinha feito um desenho pra mim	positivo
É feliz quem sonha. Más só tem sucesso quem se dispõe a pagar o preço para transformar sonho em realidade...	positivo
Anotado, vou querer ver a tatuagem depois e o cabelo tbm	neutro
Eu cometi o terrível erro de beber uma caneca de café agr de tarde	negativo
Sem querer eu descubro as coisas, esse doido mente muitooooooooo, não sei como eu fiquei com esse inútil	negativo

4. Avaliação do conjunto de dados

Antes de qualquer treinamento utilizando o conjunto de dados, realizou-se o pré-processamento do texto coletado com auxílio da biblioteca NLTK [Loper and Bird 2002]. Ele consistiu em quatro etapas: remoção de *links* por meio de expressão regular, remoção de *stopwords* em português, aplicação de *stemming* e remoção de acentuação. Após realizadas essas etapas, o texto foi vetorizado utilizado o *TFIDF Vectorizer* da biblioteca *Scikit-learn* [Pedregosa et al. 2011], formato adequado aos classificadores.

Com o texto pré-processado, foi utilizada validação cruzada com *5-folds* para avaliar o resultado dos classificadores *LogisticRegression*, *MultinomialNB*, *SGDClassifier*, *LinearSVC* e o *NuSVC*, implementados pela biblioteca *Scikit-learn*. Nela, 80% do conjunto de dados (2.229 exemplos) foi utilizado para treinamento do classificador, e 20% para teste (558 exemplos). Além disso, para fins de comparação, os exemplos do conjunto de testes também foram classificados usando o SenticNet. Nesse caso, para obter o sentimento de um *tweet*, foi calculada a média das polaridades das palavras do *tweet* existentes no SenticNet. Com o resultado em mãos, a classificação foi atribuída seguindo as regras:

- Se a média das polaridades for maior que 0,1, o *tweet* tem sentimento positivo;
- Se a média das polaridades for menor que -0,1, o *tweet* tem sentimento negativo;
- Se a média das polaridades estiver entre -0,1 e 0,1, o *tweet* tem sentimento neutro.

A Tabela 3 traz os resultados da validação cruzada e da classificação utilizando o SenticNet. Na primeira coluna tem-se a técnica de classificação e na segunda a média da acurácia dos 5 *folds*.

Tabela 3. Resultados das classificações

Classificador	Acurácia
LogisticRegression	0,4503
MultinomialNB	0,4415
SGDClassifier	0,4250
LinearSVC	0,4365
NuSVC	0,4216
SenticNet	0,3523

Analisando os resultados, percebe-se que a maior acurácia foi do classificador *LogisticRegression*, alcançando o valor 0,4503. Entretanto, a maioria dos classificadores obteve resultados semelhantes entre 0,42 e 0,45. Assim, considerando que uma classificação

aleatória de 3 classes (positivo, negativo e neutro) tende a 0,33 de acurácia, então todos os classificadores ficaram acima dessa taxa base. Em relação ao SenticNet, a acurácia foi de 0,3523, só um pouco acima da taxa base de 0,33, mostrando um desempenho ruim dessa abordagem para classificação.

A presença de linguagem informal nos *tweets*, com gírias e erros de ortografia prejudica o desempenho da classificação usando lexicons como o SenticNet, que analisa conceitos preestabelecidos. Dessa forma, no contexto de redes sociais, classificadores de aprendizagem de máquina supervisionada treinados com os dados coletados das próprias redes sociais tem o potencial de conseguir um desempenho melhor que lexicons, sendo mais adequados para problemas nessa área.

5. Conclusões

Neste trabalho foi desenvolvido um conjunto de dados para análise de sentimento na língua portuguesa utilizando mensagens do Twitter. Esse conjunto de dados tem potencial para servir de base para novas pesquisas e aplicações em análise de sentimentos no nosso idioma. Para isso, coletaram-se mensagens compartilhadas no Twitter que foram classificadas manualmente por voluntários. O conjunto de dados final possui 2.787 mensagens, sendo 888 positivas, 881 negativas e 1.018 neutras, e está disponível publicamente.

Foram realizados testes com classificadores de aprendizagem de máquina treinados com o conjunto de dados desenvolvido. A maioria dos classificadores testados obtiveram resultados semelhantes, sendo 0,4503 o máximo de acurácia conseguida pelo *LogisticRegression*. Ao comparar com o SenticNet, todas as abordagens de aprendizagem de máquina obtiveram acurácias superiores.

Como foram coletadas mais mensagens do que as classificadas pelos voluntários, como trabalhos futuros, pretende-se classificar mais *tweets*, para aumentar o tamanho do conjunto de dados. Além disso, testes mais rigorosos com classificadores de aprendizagem precisam ser realizados, incluindo técnicas como redes neurais.

6. Agradecimentos

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e a Universidade Federal da Paraíba pelo auxílio financeiro através do Programa Institucional de Bolsas de Iniciação Científica (PIBIC).

Referências

Brum, H. B. and Nunes, M. d. G. V. (2017). Building a sentiment corpus of tweets in brazilian portuguese. *arXiv preprint arXiv:1712.08917*.

Cambria, E., Poria, S., Hazarika, D., and Kwok, K. (2018). Senticnet 5: Discovering conceptual primitives for sentiment analysis by means of context embeddings. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Corrêa, E. A., Marinho, V. Q., dos Santos, L. B., Bertaglia, T. F. C., Treviso, M. V., and Brum, H. B. (2017). Pelesent: Cross-domain polarity classification using distant supervision. In *2017 Brazilian Conference on Intelligent Systems (BRACIS)*, pages 49–54. IEEE.

de Arruda, G. D., Roman, N. T., and Monteiro, A. M. (2015). An annotated corpus for sentiment analysis in political news. In *Proceedings of the 10th Brazilian Symposium in Information and Human Language Technology*, pages 101–110.

Dias-Da-Silva, B. C. and de Moraes, H. R. (2003). A construção de um thesaurus eletrônico para o português do brasil. *ALFA: Revista de Linguística*, 47(2).

Domo (2019). Data never sleeps 6. https://www.domo.com/learn/data-never-sleeps-6. Acessado em: 18-05-2019.

Dosciatti, M. M., Ferreira, L. P. C., and Paraiso, E. C. (2015). Anotando um corpus de notícias para a análise de sentimentos: um relato de experiência (annotating a corpus of news for sentiment analysis: An experience report). In *Proceedings of the 10th Brazilian Symposium in Information and Human Language Technology*, pages 121–130.

Gamon, M. (2004). Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis. In *Proceedings of the 20th international conference on Computational Linguistics*, pages 611–617. Association for Computational Linguistics.

Liu, B., Hu, M., and Cheng, J. (2005). Opinion observer: analyzing and comparing opinions on the web. In *Proceedings of the 14th international conference on World Wide Web*, pages 342–351. ACM.

Liu, B. and Zhang, L. (2012). A survey of opinion mining and sentiment analysis. In *Mining text data*, pages 415–463. Springer.

Loper, E. and Bird, S. (2002). Nltk: The natural language toolkit. In *In Proceedings of the ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics. Philadelphia: Association for Computational Linguistics*.

Moraes, S. M., Manssour, I. H., and Silveira, M. S. (2015). 7x1pt: um corpus extraído do twitter para análise de sentimentos em língua portuguesa. In *Anais do X Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 21–25. SBC.

Pang, B. and Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.