Emoções em português do Brasil: um conjunto de dados e resultados de base

Gabriel Nascimento[1], **Fellipe Duarte**[2], **Gustavo Paiva Guedes**[1]

[1]CEFET/RJ - Centro Federal de Educação Tecnológica Celso Suckow da Fonseca
Av. Maracanã, 229 - Rio de Janeiro - RJ - Brasil.

[2]UFRRJ - Universidade Federal Rural do Rio de Janeiro
Av. Gov. Roberto Silveira, s/n - Moquetá - Nova Iguaçu - Brasil

gabriel.nascimento@eic.cefet-rj.br, duartefellipe@ufrrj.br,
gustavo.guedes@cefet-rj.br

Abstract. *This paper presents a new dataset for sentiment analysis in Brazilian Portuguese. The texts were extracted from a Brazilian social network named Meu Querido Diário. In this social network, users often share feelings and emotions associated with everyday life. The main difference of this data set is that, in this social network, the user himself can inform the emotion associated with his entry. Preliminary experiments were performed with some classification models, creating the first baseline results. The model that obtained the best result was the SVM with linear kernel using bigrams.*

Resumo. *Este artigo apresenta um novo conjunto de dados para análise de sentimentos em português do Brasil. Os textos foram extraídos de uma rede social brasileira denominada Meu Querido Diário. Nessa rede social, os usuários frequentemente compartilham sentimentos e emoções associados ao dia-a-dia. O principal diferencial deste conjunto de dados é que, nessa rede social, o próprio usuário pode informar a emoção associada à sua entrada. Foram realizados experimentos preliminares com alguns modelos de classificação, criando os primeiros resultados de base. O modelo que obteve melhor resultado foi o SVM com kernel linear utilizando bigramas.*

1. Introdução

Há um vasto número de indivíduos que compartilham suas opiniões/sentimentos sobre diversos assuntos (e.g., produtos, pessoas, notícias) em redes sociais, *blogs*, *microblogs* e sites de *reviews* [Rosenthal et al., 2015]. Essa grande quantidade de dados tem despertado, cada vez mais, o interesse de pesquisadores em Análise de Sentimentos (AS), um importante campo da área de Processamento de Linguagem Natural (PLN) [Pang et al., 2008].

O objetivo da AS é analisar opiniões, atitudes, emoções, sentimentos e avaliações expressas por usuários em textos [Liu, 2012]. Logo, muitas empresas investem em pesquisas nessa área, seja para analisar informações políticas ou para investigar a percepção acerca de um produto [Cambria et al., 2013]. Dentre as tarefas existentes na AS, destaca-se a classificação de polaridade, que consiste em classificar um texto como *um* entre *dois* sentimentos opostos [Cambria et al., 2013].

A detecção de sentimentos em textos permite que os pesquisadores adquiram informações valiosas em grande escala [Rosenthal et al., 2015]. Isso reduz o tempo e aumenta a quantidade de informação que os pesquisadores podem absorver, visto que a detecção ocorre de maneira automática. No entanto, é importante destacar que a maior parte dos estudos em classificação de polaridade se concentra em textos na língua inglesa [Wiegand et al., 2010]. Dessa maneira, ainda existe uma carência de estudos concentrados no Português do Brasil (PB) [Guedes et al., 2016].

Nesse panorama, a principal contribuição desse trabalho consiste em criar um conjunto de dados de emoções em PB. Para isso, foram extraídas entradas de uma rede social brasileira denominada Meu Querido Diário (MQD)[1]. Nessa rede, os usuários podem compartilhar sentimentos/emoções sobre seus dias. Além disso, podem associar a emoção que estão sentindo no momento em que estão escrevendo suas entradas. Não foram encontrados conjuntos de dados semelhantes na literatura para o PB. Após a criação do conjunto de dados, foi produzido um resultado de base (*baseline*) nesse conjunto de dados, aplicando os algoritmos-padrão utilizados na área de AS, o que pode ser utilizado em trabalhos futuros nesse tópico.

O restante deste artigo está organizado como segue. A seção 2 discute alguns trabalhos relacionados ao tema proposto. A seção 3 descreve o conjunto de dados criado. A seção 4 descreve a metodologia de avaliação e a seção 5 apresenta a conclusão.

2. Trabalhos Relacionados

Existem na literatura alguns trabalhos que apresentam novos conjuntos de dados de redes sociais para realização de análise de sentimentos. Podemos destacar o trabalho realizado por Saif et al. [2013], que apresentou um novo conjunto de dados baseado no Twitter. Este conjunto de dados foi nomeado *STS-Gold* e nele, todos os *tweets* foram rotulados com sentimentos: positivo, negativo, neutro, misto, outro. Os *tweets* que foram rotulados como "misto" são aqueles que possuem mais de um tipo de sentimento envolvido enquanto os rotulados com "outro" são difíceis de classificar.

Brum and Nunes [2017] construíram um conjunto de dados em PB, anotado a partir do Twitter, com base nas hashtags referentes a programas de televisão brasileiros. Cada *tweet* foi rotulado manualmente com as classes de sentimento positivo, negativo ou neutro. Contudo, apesar do conjunto de dados apresentar os *ids* dos Tweets, os dados não foram disponibilizados em sua integridade por conta da política de privacidade do Twitter, porém foram disponibilizados os *ids* dos Tweets. Além disso, os conjuntos de dados anotados apresentam uma limitação em relação a quantidade de texto disponível por Tweet visto que um Tweet pode ter até 280 caracteres[2].

De Pelle et al. de Pelle and Moreira [2017] criaram um conjunto de dados em PB, anotado para tarefas de detecção de discurso de ódio com base em comentários de usuários da internet. A coleta dos dados foi efetuada coletando comentários do site de notícias *g1.globo.com* onde, foi realizado um processo de julgamento utilizando um sistema[3], com três juízes, para rotular a classe dos comentários. Cada comentário foi classificado como ofensivo ou não, e caso classificado como ofensivo, poderia possuir de uma a seis

[1] http://www.meuqueridodiario.com.br

[2] eram 140 antes de novembro de 2017

[3] http://inf.ufrgs.br/~rppelle/hatedetector/

classes: racismo, sexismo, homofobia, xenofobia, intolerância religiosa ou xingamento. Os experimentos realizados consistiram em classificar automaticamente se um comentário contém discurso de ódio ou não.

Embora existam alguns conjuntos de dados textuais rotulados em PB (e.g., postagens no Twitter [Nascimento et al., 2012], notícias [de Pelle and Moreira, 2017]), o conjunto de dados aqui proposto se diferencia por não limitar a quantidade de caracteres permitidos (no caso do Twitter, são até 280 caracteres) e pelo próprio usuário ser o rotulador do seu texto, expressando, de forma voluntária, seu sentimento ao escrever.

3. Conjunto de dados para Análise de Sentimentos em PT-BR

Conforme mencionado nas seções anteriores, a maioria dos estudos no âmbito da AS estão concentrados em textos em inglês, o que motivou o presente trabalho a propor um conjunto de dados em PB. Para isto, foram extraídos dados textuais das entradas dos usuários da rede social brasileira MQD. Nesta rede social, os usuários compartilham seus sentimentos e emoções relacionados aos seus dias. O diferencial deste conjunto de dados é destacado pelo fato de o próprio usuário, que escreveu a entrada, ter realizado a classificação da emoção. Isso ocorre porque os usuários podem associar às suas entradas *uma* entre seis emoções (i.e., felicidade, tristeza, raiva, medo, nojo e surpresa).

No momento da extração dos dados[4], o MQD possuía 69.452 usuários cadastrados e 191.042 entradas. Do total de usuários, 32.546 possuem 1 ou mais entradas das quais foram selecionadas as entradas dos usuários que decidiram associar uma emoção ao realizar a escrita. Desta maneira, foram selecionadas 79.523 entradas e o número de entradas associada a cada emoção é distribuído da seguinte maneira: felicidade, 32.672; tristeza, 27.642; raiva, 6.323; medo, 6.112; surpresa, 5.262; nojo, 1.512;

O conjunto de dados apresentado foi denominado `MQDEmotion2018`[5]. Ao todo foram contabilizas 18.601.010 palavras e um vocabulário de 265.741 termos (palavras distintas) no conjunto de dados. Vale destacar que o `MQDEmotion2018` é um conjunto de dados para AS em PB, extraído diretamente de uma atividade cotidiana e, portanto, um dos maiores desafios desse conjunto de dados é a quantidade de palavras escritas com grafia incorreta, pois gera um grande número de palavras com apenas uma ocorrência.

4. Metodologia de avaliação

Os modelos de pré-processamento e seleção de atributo utilizados nos experimentos são apresentados nesta seção que é dividida em duas subseções: a Subseção 4.1 descreve a metodologia empregada para pré-processar o conjunto de dados utilizado nos experimentos, ou seja, o `MQDEmotion2018`. Em seguida, a Subseção 4.2 apresenta a metodologia adotada para a execução dos algoritmos de classificação.

4.1. Pré-processamento

O pré-processamento dos dados foi dividido em três fases. A primeira fase consistiu em selecionar um subconjunto dos dados do `MQDEmotion2018`. Isso ocorreu devido à tarefa proposta neste artigo, ou seja, a classificação de polaridade. Desta maneira, para os experimentos deste trabalho, foram utilizadas apenas entradas que representam as emoções *felicidade* e *tristeza*, o que corresponde a 60.314 entradas do

[4] A extração foi realizada em 10 de janeiro de 2018.

[5] Disponível em https://github.com/LaCAfe/MQDEmotion2018

`MQDEmotion2018`. Do total de entradas, 32.672 são rotuladas com a emoção *felicidade* e 27.642 com a emoção *tristeza*. Com isto, o número de palavras totalizou $13.926.356$ e o total de termos únicos, 220.540.

Em seguida, na segunda fase do pré-processamento, foram removidas as palavras com *uma* ocorrência, dado que elas podem piorar a tarefa de classificação Zhu and Chen [2005]. Desta maneira, o número de palavras totalizou $13.796.872$ e o número de termos únicos correspondeu a 91.056. Por fim, na terceira fase do pré-processamento, os textos dos documentos foram, inicialmente, convertidos para letras minúsculas. Em seguida, os documentos foram representados como vetores TF-IDF (*Term frequency - inverse document frequency*), uma vez que esta é a representação vetorial mais bem sucedida para na tarefa de categorização de textos Salton et al. [1975]. Estes são os dados utilizados no decorrer do presente artigo. Para melhor entendimento das demais seções, o conjunto de dados resultante das três fases de pré-processamento é denominado `MQDEmotion2018ft`.

4.2. Classificadores

Os experimentos utilizam o classificador Multinomial Naïve Bayes (MNB), por ser considerado o *baseline* por alguns trabalhos [Arias et al., 2013; Hassan and Mahmood, 2017]. Também utiliza o *Linear Support Vetor Machines* (LSVC), por ter apresentado bons resultados em tarefas de AS [Arias et al., 2013; Hassan and Mahmood, 2017]. Desta maneira, esses classificadores são empregados para avaliar o `MQDEmotion2018ft`.

O classificador *Naïve Bayes* supõe que todos os atributos extraídos das amostras fornecidas são independentes, dada uma hipótese em um contexto de classificação [McCallum et al., 1998]. Os autores ainda destacam que, embora essa presunção não seja verdade para a maioria dos problemas reais, o classificador *Naïve Bayes* tende a ter um bom desempenho. O *Multinomial Naïve Bayes* (MNB) é uma implementação do classificador *Naïve Bayes* e assume que os dados seguem uma distribuição multinomial, utilizando como informação o número de vezes que o termo ocorre em cada documento [McCallum et al., 1998].

O *Support Vector Machine* (SVM) tem como objetivo encontrar um hiperplano ótimo entre os dados em um espaço vetorial, dividindo-os em classes [Cortes and Vapnik, 1995]. Este algoritmo tem apresentado grande sucesso na tarefa de classificação em texto por ser efetivo em espaços de alta dimensionalidade e em tarefas que a dimensionalidade dos dados é maior que o número de amostras disponíveis [Forman, 2007]. Logo, utilizando os documentos de uma classe como pontos em um espaço n-dimensional, um classificador LSVC separa linearmente o espaço em que os pontos das classes devem pertencer.

4.3. Análise Experimental

Para realizar os experimentos, o trabalho foi organizado como segue. Os classificadores foram treinados e avaliados a partir da estratégia de validação cruzada de 10 partições (*10-fold cross-validation*).

Conforme procedido em Hassan and Mahmood [2017], além dos classificadores MNB e SVM linear (LSVC), foram utilizadas duas representações de N-gramas: unigrama (1-grama) e bigramas (2-gramas). Essas representações são encontradas na Tabela 1. A tarefa de classificação do MNB com 1-grama e 2-gramas foi denominada MNB^1

e MNB^2, respectivamente. Analogamente, a tarefa de classificação do LSVC com 1-grama e 2-gramas foi denominada $LSVC^1$ e $LSVC^2$. Os classificadores MNB foram avaliados com diferentes valores de *smoothing*(α), com $\alpha = 0$, $\alpha = 0.5$ e $\alpha = 1$.

Atributos	MNB	LSVC
1-grama	MNB^1	$LSVC^1$
2-grama	MNB^2	$LSVC^2$

Tabela 1. Rótulos de identificação da combinação entre a representação dos atributos e dos classificadores.

A Tabela 2 exibe a média dos resultados de acurácia, *F1 score* e seus respectivos desvios padrões obtidos para os modelos treinados. O modelo $LSVC^2$ obteve os melhores resultados, tanto na acurácia, quanto no *F1 score*. Ademais, o uso de 2-gramas provou ser melhor na tarefa, aumentando o *F1 score* e a acurácia de todos os modelos.

Classificador	Acurácia	F1
$MNB^1(\alpha = 0)$	$0,72 \pm 0,02$	$0,69 \pm 0,02$
$MNB^1(\alpha = 0,5)$	$0,78 \pm 0,02$	$0,77 \pm 0,02$
$MNB^1(\alpha = 1)$	$0,78 \pm 0,02$	$0,77 \pm 0,02$
$LSVC^1$	$0,81 \pm 0,01$	$0,79 \pm 0,01$
$MNB^2(\alpha = 0)$	$0,74 \pm 0,02$	$0,70 \pm 0,02$
$MNB^2(\alpha = 0,5)$	$0,79 \pm 0,01$	$0,78 \pm 0,02$
$MNB^2(\alpha = 1)$	$0,79 \pm 0,02$	$0,78 \pm 0,02$
$LSVC^2$	**0,82** $\pm 0,01$	**0,80** $\pm 0,01$

Tabela 2. Acurácia e *F1 score* obtidos com os modelos MNB e LSVC na tarefa de AS.

5. Conclusão

Este artigo apresenta como principal colaboração um novo conjunto de dados para análise de sentimentos. Esse conjunto de dados foi retirado de uma rede social brasileira denominada Meu Querido Diário. O principal diferencial desse conjunto de dados é que o próprio usuário fornece, de forma espontânea, a emoção associada ao seu texto no momento em que o estava escrevendo.

Os *baseline* escolhidos foram os algoritmos *Multinomial Naive Bayes* e *Linear Support Vector Classifier* para realizar as tarefas de classificação. De forma geral, os classificadores que utilizaram 2-gramas apresentaram melhorias na acurácia e no *F1 score* de todos os classificadores avaliados onde o classificador que obteve melhor desempenho foi o *Linear Support Vector Classifier* com 2-gramas assim como apresentado em Hassan and Mahmood [2017] para a tarefa de AS em inglês.

Em trabalhos futuros serão estudados outros modelos para realizar a classificação neste conjunto de dados. Com o objetivo de melhorar o desempenho das tarefas de análise de sentimentos, também serão analisadas melhores técnicas de pré-processamento e outras formas de representação de atributos, como por exemplo, *word embeddings*.

Referências

Arias, M., Arratia, A., and Xuriguera, R. (2013). Forecasting with twitter data. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 5(1):8.

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., Schuller, B., Xia, Y., and Havasi, C. (2013). New avenues in opinion mining and sentiment analysis. *IEEE Intelligent Systems*, 28(2):15–21.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3):273–297.

de Pelle, R. P. and Moreira, V. P. M. (2017). Offensive comments in the brazilian web: a dataset and baseline results. In *Congresso da Sociedade Brasileira de Computação-CSBC*.

Forman, G. (2007). Feature selection for text classification. *Computational methods of feature selection*, 1944355797.

Guedes, G. P., Bezerra, E., Ferrari, L., and Duarte, F. (2016). Gender differences in the use of portuguese in social networks: Evidence from liwc. In *Proceedings of the 22nd Brazilian Symposium on Multimedia and the Web*, pages 339–342. ACM.

Hassan, A. and Mahmood, A. (2017). Deep learning for sentence classification. In *Systems, Applications and Technology Conference (LISAT), 2017 IEEE Long Island*, pages 1–5. IEEE.

Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167.

McCallum, A., Nigam, K., et al. (1998). A comparison of event models for naive bayes text classification. In *AAAI-98 workshop on learning for text categorization*, volume 752, pages 41–48. Citeseer.

Nascimento, P., Aguas, R., Lima, D., Kong, X., Osiek, B., Xexéo, G., and Souza, J. (2012). Análise de sentimento de tweets com foco em notícias. In *Brazilian Workshop on Social Network Analysis and Mining*.

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

Rosenthal, S., Nakov, P., Kiritchenko, S., Mohammad, S., Ritter, A., and Stoyanov, V. (2015). Semeval-2015 task 10: Sentiment analysis in twitter. In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, pages 451–463.

Saif, H., Fernandez, M., He, Y., and Alani, H. (2013). Evaluation datasets for twitter sentiment analysis: a survey and a new dataset, the sts-gold.

Salton, G., Wong, A., and Yang, C.-S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620.

Wiegand, M., Balahur, A., Roth, B., Klakow, D., and Montoyo, A. (2010). A survey on the role of negation in sentiment analysis. In *Proceedings of the workshop on negation and speculation in natural language processing*, pages 60–68. Association for Computational Linguistics.

Zhu, J. and Chen, W. (2005). Some studies on chinese domain knowledge dictionary and its application to text classification. In *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*.