Utilizando BERTimbau para a Classificação de Emoções em Português

Luiz Otávio Alves Hammes[1], **Larissa Astrogildo de Freitas**[1]

[1]Centro de Desenvolvimento Tecnológico (CDTec)
Universidade Federal de Pelotas (UFPel) – Pelotas, RS – Brasil

{loahammes,larissa}@inf.ufpel.edu.br

Abstract. *In this work, we propose the fine-tuning of the BERTimbau-base and BERTimbau-large models in the task of Sentence Classification with 27 types of emotions, based on the GoEmotions dataset translated into Portuguese, using automatic translation tools. We compared the results of our experiments with the one provided by the authors of the GoEmotions dataset and obtained a performance gain which we attribute to the balancing algorithm used.*

Resumo. *Neste trabalho propomos realizar o fine-tuning dos modelos BERTimbau-base e BERTimbau-large na tarefa de Classificação de 27 tipos de emoções em sentenças, baseado no dataset GoEmotions traduzido para a língua portuguesa, por meio de ferramentas de tradução automática. Comparamos os resultados de nossos experimentos com os resultados disponibilizados pelos autores do dataset GoEmotions e obtivemos um ganho de desempenho ao qual atribuímos ao algoritmo de balanceamento utilizado.*

1. Introdução

Análise de Sentimento (AS) é uma tarefa da área de Processamento de Língua Natural (PLN), que tem como objetivo extrair, processar e classificar opiniões, sentimentos, emoções e avaliações da vasta quantidade de conteúdo disponível na Web atualmente, podendo ser textos, áudios ou imagens [Liu 2012].

Ainda, podemos dividir as tarefas de AS em três níveis: (1) nível de documento, no qual um sentimento é atribuído ao documento como um todo ; (2) nível de sentença, no qual um sentimento é atribuído a cada sentença presente em um documento; (3) nível de aspecto, no qual são identificados aspectos de uma entidade[1] presentes em um documento e então um sentimento é atribuído para cada um dos aspectos encontrados.

A Classificação de Emoções (CE) em nível de sentença é uma subárea da AS, a qual tem como objetivo classificar as distintas emoções expressas em um conjunto de sentenças que, em sua grande maioria, são de caráter subjetivo. Isso acontece devido as sentenças subjetivas manifestarem emoções, que estão atreladas as percepções, os pontos de vista e o estado emocional dos seus autores [Liu 2012].

Definir o conceito de emoção é uma tarefa difícil. Isso ocorre devido as emoções estarem relacionadas a uma imensa quantidade de distintos fatores que podem fazer com que elas sejam manifestadas. Uma das principais teorias é a psicoevolutiva, a qual sugere

[1]Uma entidade é algo que pode ser nomeado, como: pessoa, objeto, produto, localização, etc.

a existência de manifestações básicas de emoções que são consequência de processos evolutivos, os trabalhos de [Ekman 2004] e de [Plutchik 2003] defendem essa ideia.

Avanços recentes na psicologia sugerem a existência de um conjunto maior de emoções, manifestadas a partir de diferentes contextos e situações. Por exemplo, o trabalho de [Cowen and Keltner 2017] identificou 27 emoções expressas por pessoas a partir de vídeos curtos e o trabalho de [Cowen and Keltner 2020] identificou 28 emoções expressas através de expressões faciais e linguagem corporal.

Tendo em vista este contexto, percebemos que é necessário ampliar a quantidade de classes de emoções dos modelos classificadores disponíveis na literatura para abranger estes estudos recentes, já que a maioria dos trabalhos utiliza como base apenas os trabalhos de [Ekman 1992] (6 emoções básicas) e de [Plutchik 1982] (8 emoções básicas).

Neste trabalho, propomos utilizar modelos de Aprendizado de Máquina (AM) para classificar 27 emoções com base no *dataset* GoEmotions [Demszky et al. 2020], traduzido para o português, com a finalidade de melhorar os resultados preliminares e ampliar a variedade de classes na tarefa de CE em língua portuguesa, disponíveis na literatura.

Este artigo está estruturado da seguinte forma: a Seção 2 apresenta os trabalhos relacionados; a Seção 3 descreve a metodologia; a Seção 4 mostra a análise dos resultados; e a Seção 5 apresenta as conclusões.

2. Trabalhos Relacionados

De acordo com [Pereira 2021], poucos são os trabalhos encontrados na literatura que focam na tarefa de CE em língua portuguesa. Dentre eles podemos citar o trabalho de [Duarte et al. 2019] e de [Dosciatti et al. 2013].

No trabalho de [Duarte et al. 2019] foram utilizados emojis para reconhecer 6 emoções básicas (felicidade, raiva, nojo, medo, tristeza e surpresa) utilizando a taxonomia de [Ekman 1992] em *tweets*. Os autores coletaram 2 milhões de *tweets* que continham pelo menos um dos emojis que segundo [Wood and Ruder 2016] são indicadores da presença das 6 emoções básicas. Em seguida realizaram um etapa de pré-processamento, na qual foram retirados *retweets, URLs*, menções de usuários, *tweets* com dois ou mais emojis que expressam emoções contraditórias e *tweets* com menos de 3 *tokens*. Os emojis foram retirados dos *tweets* e utilizados como *labels* para treinamento dos modelos classificadores. Foram treinados dois modelos: *Support Vector Machine* (SVM) e *Naive Bayes* (NB), atingindo 0,62 e 0,70 de Medida-F, respectivamente.

No trabalho de [Dosciatti et al. 2013] foi usado um corpus com 1750 notícias curtas para treinar um modelo classificador, rotuladas nas 6 emoções básicas de [Ekman 1992] mais a classe "neutro"[2]. Inicialmente, os autores coletaram notícias do site `www.globo.com`, depois realizaram uma etapa de pré-processamento, na qual retiraram acentos, caracteres especiais e *stopwords* e converteram todos os textos para letras minúsculas. Em seguida, dois anotadores classificaram os textos nas 6 emoções básicas ou na classe "neutro". O corpus foi construído mantendo uma proporção entre as 7 classes para evitar desbalanceamento, resultando em 1750 notícias, 250 para cada classe. Os autores treinaram o modelo SVM, atingindo 0,60 de Medida-F.

[2]A classe "neutro" representa um texto que não contém conteúdo emocional.

No contexto da língua inglesa, podemos citar o trabalho de [Demszky et al. 2020] que desenvolveu o *dataset* GoEmotions com taxonomia de 27 emoções, baseando-se nos trabalhos de [Cowen and Keltner 2017], [Cowen et al. 2019a], [Cowen and Keltner 2020], [Cowen et al. 2019b], contendo 54263 sentenças. Ainda, os autores disponibilizaram um modelo BERT-base [Devlin et al. 2018] com *fine-tuning* utilizando o *dataset* GoEmotions. Mais detalhes sobre o *dataset* GoEmotions serão discutidos na Seção 3, já que ele será utilizado como base para nosso trabalho.

3. Metodologia

O processo de execução desse trabalho foi divido em três etapas principais. A concepção do modelo, *fine-tuning* e avaliação dos resultados obtidos.

Em nossos experimentos utilizamos a versão filtrada do *dataset* GoEmotions, que contém 54263 sentenças manualmente anotadas em 28 classes (27 emoções + neutro) retiradas do fórum Reddit na língua inglesa. O processo de filtragem aplicado por [Demszky et al. 2020] é realizado em duas etapas, primeiro são retirados todos os *labels* que foram selecionados por apenas um anotador e depois são mantidas apenas as sentenças com pelo menos um *label*. Manteve-se a proporção original: 80% treinamento, 10% validação e 10% teste.

Como o objetivo deste trabalho é CE em sentenças escritas em português, traduzimos o *dataset* GoEmotions com o auxílio da biblioteca itranslate [3] que facilita a utilização da API (*Application Programming Interface*) do Google Translate[4].

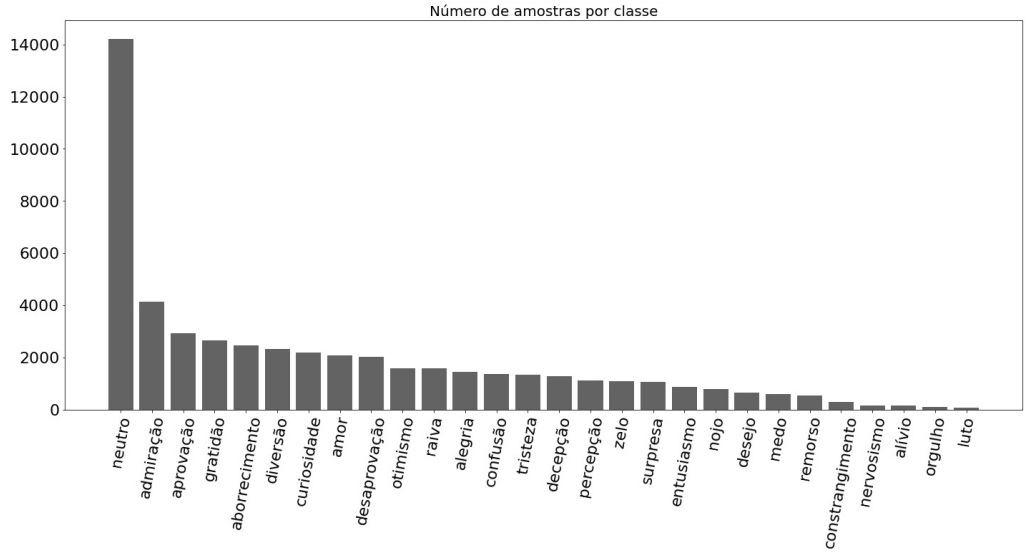

Figura 1. Distribuição das amostras por classe de emoção no *dataset* GoEmotions. Fonte: Própria

Sobre a distribuição das quantidades de exemplos no *dataset* GoEmotions (Figure 1), é possível notar um grande desbalanceamento. A classe mais frequente "neutro", tem aproximadamente 184 vezes mais amostras do que a classe "luto" que é a menos frequente. Caso nada for feito para amenizar este problema, o modelo classificador desen-

[3]https://github.com/ffreemt/google-itranslate

[4]https://translate.google.com.br/

volverá um viés, que resultará em uma habilidade preditiva ruim nas classes com poucos exemplos [Zheng and Jin 2020]. Para contornar esse problema utilizamos um método de balanceamento que será detalhado a seguir.

3.1. Modelo

O modelo BERT (*Bidirectional Encoder Representations from Transformers*) [Devlin et al. 2018], mais especificamente a versão pré-treinada para o português BERTimbau [Souza et al. 2020], foi escolhido para esta tarefa, pois modelos baseados na arquitetura *Transformers* [Vaswani et al. 2017], como o BERT, têm apresentado desempenho estado da arte nas mais diversas tarefas de PLN [Gillioz et al. 2020].

Como o *dataset* GoEmotions é *multi-label*, ou seja, pode ter um ou mais *labels* para cada exemplo, codificamos todos os *labels* atribuídos a cada sentença em um vetor *one-hot* e transformamos a tarefa em um problema de classificação binária para cada classe do modelo. Foi adicionada uma camada linear sobre o *pooled output* do modelo com um tamanho de saída igual a 28, equivalente ao número de classes do *dataset*. Utilizamos *Sigmoid* como função de ativação e *Binary Cross Entropy* como função de *loss*.

Utilizamos o método *Class Balanced Loss* (**CB**) para contornar o problema de desbalanceamento do *dataset* [Cui et al. 2019] . Neste método são calculados pesos para a função de *loss* ($\mathcal{L}$) utilizada no modelo, com base no número efetivo de amostras para cada classe. Um hiperparâmetro $\beta = 0.999$, também é utilizado para o cálculo destes pesos. Uma vez que, boa parte dos experimentos de [Cui et al. 2019] obteve um bom desempenho com este valor. A nova função que computa o *loss* do modelo é obtida por meio da seguinte equação:

$$\mathbf{CB}(p, y) = \frac{1 - \beta}{1 - \beta^{n_y}} \mathcal{L}(p, y), \tag{1}$$

onde p são as probabilidades fornecidas pelo modelo e n_y é o número de amostras presentes na classe y.

3.2. Fine-tuning

Fine-tuning é uma técnica de *transfer learning* na qual partindo de um modelo pré-treinado em uma tarefa de amplo domínio, os parâmetros desse modelo são ajustados para uma tarefa específica. Neste trabalho realizaremos o *fine-tuning* do BERTimbau-base e BERTimbau-large para a tarefa de CE no nível de sentenças. Os hiperparâmetros utilizados no *fine-tuning* do modelo estão listados na Tabela 1. Por limitação do hardware disponível, foi feito o *fine-tuning* do modelo BERTimbau-base em GPU e BERTimbau-large em CPU.

4. Análise dos Resultados

Para avaliar o desempenho do modelo foram utilizadas as métricas Precisão (P), Sensibilidade (S) e Medida-F (F1), mesmas métricas utilizadas pelos autores do *dataset* GoEmotions [Demszky et al. 2020], os quais utilizam em seus experimentos o modelo BERT-base para o inglês [Devlin et al. 2018]. Selecionamos apenas as predições do modelo com um *score* igual ou superior a 0,3. A Tabela 2 apresenta a comparação entre os resultados reportados por [Demszky et al. 2020] e pelo modelo BERTimbau-base e BERTimbau-large com o método de balanceamento, na tarefa de CE de 27 emoções no nível de sentença.

Hiperparâmetros	BERTimbau-base	BERTimbau-large
epochs	4	4
batch size	16	32
β (beta)	0.999	0.999
maximum sequence length	128	128
optimizer	*AdamW*	*AdamW*
learning rate	2e-5	2e-5
learning rate scheduler	*linear with warmup*	*linear with warmup*
warmup proportion	0.2	0.2

Tabela 1. Hiperparâmetros utilizados no *fine-tuning*.

Emoções	BERT-base EN [Devlin et al. 2018]			BERTimbau-base PT [Souza et al. 2020] Balanceado			BERTimbau-large PT [Souza et al. 2020] Balanceado		
	P	S	F1	P	S	F1	P	S	F1
admiração	0,53	0,83	0,65	0,58	0,75	0,66	0,60	0,75	0,67
diversão	0,70	0,94	0,80	0,76	0,89	0,82	0,75	0,91	0,82
raiva	0,36	0,66	0,47	0,37	0,46	0,41	0,40	0,49	0,44
aborrecimento	0,24	0,63	0,34	0,37	0,33	0,35	0,35	0,38	0,36
aprovação	0,26	0,57	0,36	0,40	0,40	0,40	0,42	0,40	0,41
zelo	0,30	0,56	0,39	0,34	0,44	0,39	0,34	0,47	0,39
confusão	0,24	0,76	0,37	0,33	0,56	0,41	0,35	0,58	0,44
curiosidade	0,40	0,84	0,54	0,44	0,77	0,56	0,45	0,80	0,57
desejo	0,43	0,59	0,49	0,53	0,55	0,54	0,60	0,58	0,59
decepção	0,19	0,52	0,28	0,28	0,20	0,23	0,28	0,25	0,26
desaprovação	0,29	0,61	0,39	0,37	0,43	0,40	0,39	0,45	0,42
nojo	0,34	0,66	0,45	0,46	0,42	0,44	0,46	0,47	0,46
constrangimento	0,39	0,49	0,43	0,39	0,38	0,38	0,39	0,35	0,37
entusiasmo	0,26	0,52	0,34	0,36	0,43	0,39	0,35	0,50	0,41
medo	0,46	0,85	0,60	0,54	0,73	0,62	0,57	0,77	0,65
gratidão	0,79	0,95	0,86	0,88	0,91	0,90	0,88	0,92	0,90
luto	0,00	0,00	**0,00**	0,13	0,67	**0,22**	0,17	0,50	0,25
alegria	0,39	0,73	0,51	0,51	0,57	0,54	0,51	0,60	0,55
amor	0,68	0,92	0,78	0,72	0,87	0,79	0,70	0,85	0,77
nervosismo	0,28	0,48	0,35	0,29	0,48	0,36	0,24	0,43	0,31
neutro	0,56	0,84	0,68	0,64	0,70	0,67	0,64	0,70	0,67
otimismo	0,41	0,69	0,51	0,52	0,56	0,54	0,53	0,53	0,53
orgulho	0,67	0,25	0,36	0,39	0,44	0,41	0,36	0,50	0,42
percepção	0,16	0,29	0,21	0,37	0,12	0,18	0,34	0,21	0,26
alívio	0,50	0,09	0,15	0,14	0,27	0,18	0,18	0,55	0,27
remorso	0,53	0,88	0,66	0,51	0,88	0,64	0,54	0,88	0,67
tristeza	0,38	0,71	0,49	0,47	0,54	0,50	0,48	0,57	0,52
surpresa	0,40	0,66	0,50	0,49	0,60	0,54	0,46	0,58	0,51
média macro	0,40	0,63	**0,46**	0,45	0,55	**0,48**	0,45	0,57	0,50

Tabela 2. Comparativo entre os modelos para inglês e português.

Podemos observar que ao utilizarmos a média macro da métrica Medida-F como parâmetro de avaliação do desempenho geral dos modelos analisados, o modelo para o português (BERTimbau-base), que tem a mesma quantidade de parâmetros do modelo utilizado para o inglês (BERT-base), apresentou desempenho superior. É possível atribuir isso ao método de balanceamento utilizado (*Class Balanced Loss*), já que em outros experimentos realizados não se utilizou o método de balanceamento e o desempenho das classes com a menor quantidade de exemplos foi consideravelmente prejudicado (Tabela 3). O maior ganho foi na classe "luto" que contém a menor quantidade de exemplos no *dataset*, passando de 0,00 para 0,22 na Medida-F. Um ponto importante é que o modelo obtém ótimos resultados nas classes "gratidão", "diversão" e "amor", acreditamos que isso ocorre devido a existência de palavras ou de expressões que contribuem fortemente para a identificação dessas emoções como "obrigado", "te amo/eu amo", ou gírias como "*lol, lmao, lmfao*" que estão presentes na maioria das sentenças relacionadas a essas emoções e dificilmente ocorrem em exemplos rotulados com outras emoções.

Outro ponto importante que deve ser levado em consideração é a qualidade de tradução do *dataset*. Por exemplo, a frase: "It's a better option because it's my life and none of your business? Lmfao, who are you", presente na base de treinamento, foi traduzida para: "É uma opção melhor porque é minha vida e nenhum da sua empresa? Lmfao, quem é você" , é possível perceber que na expressão idiomática "none of your business" foi realizada uma tradução literal dos seus termos, uma melhor tradução seria utilizar outra expressão idiomática com sentido equivalente para a língua portuguesa, como: "não é da sua conta". Esse tipo de problema leva a uma deterioração, ou completa perda, do sentido completo expresso pelas sentenças, podendo ser o suficiente para que a frase traduzida expresse uma emoção diferente, ou até mesmo, nenhuma emoção.

5. Conclusões

Por fim, podemos concluir que é necessária a criação de um *dataset* anotado, com uma boa variedade de emoções, no nível de sentença a ser utilizado na tarefa de CE em português, visto que, não encontramos esse tipo de recurso disponível na literatura.

Ainda, percebemos que a utilização do método de balanceamento e o modelo em português BERTimbau-base no *dataset* GoEmotions traduzido obteve melhores resultados se compararmos com o modelo em inglês no *dataset* GoEmotions original.

Como trabalho futuro pretendemos investigar o uso de diferentes ferramentas de tradução automática capazes de identificar e traduzir corretamente expressões idiomáticas, pois a qualidade da ferramenta de tradução utilizada deve impactar significativamente o desempenho dos modelos. Além disso, pretendemos construir um novo *dataset* anotado na mesma taxonomia de emoções empregada na construção do *dataset* GoEmotions, mas com sentenças provenientes de falantes da língua portuguesa. Pretendemos utilizar esses dados para verificar quanto desempenho é retido pelos modelos treinados com dados traduzidos, quando avaliados em sentenças originalmente em português.

Todos os códigos necessários para *download*, tradução do *dataset* e *fine-tuning* dos modelos utilizados estão disponíveis no GitHub e podem ser acessados através do link `https://github.com/Luzo0/GoEmotions_portuguese`.

Emoções	BERTimbau-base PT [Souza et al. 2020] Não Balanceado		
	P	S	F1
admiração	0,59	0,74	0,66
diversão	0,76	0,87	0,81
raiva	0,40	0,45	0,42
aborrecimento	0,36	0,32	0,34
aprovação	0,39	0,41	0,40
zelo	0,40	0,41	0,41
confusão	0,44	0,51	0,47
curiosidade	0,48	0,73	0,58
desejo	0,60	0,42	0,50
decepção	0,36	0,20	0,26
desaprovação	0,39	0,42	0,41
desgosto	0,52	0,41	0,46
constrangimento	0,00	0,00	**0,00**
entusiasmo	0,44	0,36	0,40
medo	0,55	0,76	0,64
gratidão	0,91	0,90	0,90
luto	0,00	0,00	**0,00**
alegria	0,52	0,56	0,54
amor	0,70	0,83	0,76
nervosismo	0,00	0,00	**0,00**
otimismo	0,57	0,57	0,57
orgulho	0,00	0,00	**0,00**
percepção	0,41	0,12	0,19
alívio	0,00	0,00	**0,00**
remorso	0,53	0,84	0,65
tristeza	0,44	0,56	0,49
surpresa	0,51	0,54	0,53
neutro	0,64	0,72	0,67
média macro	0,43	0,45	**0,43**

Tabela 3. Resultados sem o método de balanceamento.

Referências

Cowen, A., Sauter, D., Tracy, J. L., and Keltner, D. (2019a). Mapping the passions: Toward a high-dimensional taxonomy of emotional experience and expression. *Psychological Science in the Public Interest*, 20(1):69–90.

Cowen, A. S., Elfenbein, H. A., Laukka, P., and Keltner, D. (2019b). Mapping 24 emotions conveyed by brief human vocalization. *American Psychologist*, 74(6):698–712.

Cowen, A. S. and Keltner, D. (2017). Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *National Academy of Sciences*, 114(38):7900–7909.

Cowen, A. S. and Keltner, D. (2020). What the face displays: Mapping 28 emotions conveyed by naturalistic expression. *American Psychologist*, 75(3):349.

Cui, Y., Jia, M., Lin, T.-Y., Song, Y., and Belongie, S. (2019). Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9268–9277.

Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A., Nemade, G., and Ravi, S. (2020). Goemotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040—-4054. ACL.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dosciatti, M. M., Ferreira, L., and Paraiso, E. (2013). Identificando emoçoes em textos em português do brasil usando máquina de vetores de suporte em soluçao multiclasse. *ENIAC-Encontro Nacional de Inteligência Artificial e Computacional. Fortaleza, Brasil*.

Duarte, L., Macedo, L., and Oliveira, H. G. (2019). Exploring emojis for emotion recognition in portuguese text. In *Proceedings of the EPIA Conference on Artificial Intelligence*, pages 719–730. Springer.

Ekman, P. (1992). An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200.

Ekman, P. (2004). Emotions revealed. *Bmj*, 328(Suppl S5).

Gillioz, A., Casas, J., Mugellini, E., and Abou Khaled, O. (2020). Overview of the transformer-based models for nlp tasks. In *Proceedings of the 15th Conference on Computer Science and Information Systems*, pages 179–183. IEEE.

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

Pereira, D. A. (2021). A survey of sentiment analysis in the portuguese language. *Artificial Intelligence Review*, 54(2):1087–1115.

Plutchik, R. (1982). A psychoevolutionary theory of emotions. *Social Science Information*, 21(4-5):529–553.

Plutchik, R. (2003). *Emotions and life: Perspectives from psychology, biology, and evolution.* American Psychological Association.

Souza, F., Nogueira, R., and Lotufo, R. (2020). Bertimbau: pretrained bert models for brazilian portuguese. In *Proceedings of the Brazilian Conference on Intelligent Systems*, pages 403–417. Springer.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Proceedings of the Advances in neural information processing systems*, pages 5998–6008.

Wood, I. and Ruder, S. (2016). Emoji as emotion tags for tweets. In *Proceedings of the Emotion and Sentiment Analysis Workshop LREC2016, Portorož, Slovenia*, pages 76–79.

Zheng, W. and Jin, M. (2020). The effects of class imbalance and training data size on classifier learning: an empirical study. *SN Computer Science*, 1(2):1–13.