

Validação e construção de um dicionário léxico para auxiliar a análise de sentimentos em repositórios de projetos de software

Hiolanda Menezes, Gláucia Boechat, Joselito Mota Jr, Ivan Machado

¹Universidade Federal da Bahia (UFBA) – Salvador – BA – Brazil

{hiolanda.murielle, glaucya.boechat, joselito.mota, ivan.machado}@ufba.br

Abstract. *Sentiment analysis makes inference about polarities in words that can represent possible emotions. The assertiveness of this classification is important for the results reliability. For this reason, this article investigates, validates and builds the lexicon dictionary, in the context of Software Engineering, using words, emoticons and idiomatic expressions from the SentiStrength-SE tool. An experiment online with 559 questions answered from 48 participants in the Computing area was performed to validate lexical terms agreement from the dictionary. At the end of the data collection, the terms were gathered for validation using a Stack Overflow database to find the results on accuracy, precision, recall and F_1 -score of the new dictionary. The new lexical dictionary has 79% Accuracy and Precision, with 78% Recall and $f1$ -score with a smaller polarity interval than the original dictionary.*

Resumo. *A análise de sentimentos faz inferência sobre polaridades em palavras que podem representar possíveis emoções. A assertividade dessa classificação é importante para a confiabilidade do resultado esperado. Por esta razão, este trabalho busca investigar, validar e construir um dicionário léxico, no contexto de Engenharia de Software, utilizando como base as palavras, emoticons e expressões idiomáticas da ferramenta SentiStrength-SE. Um experimento com 559 questões respondidas por 48 participantes da área de Computação foi realizado para validação da concordância dos termos léxicos do dicionário. Ao final da coleta dos dados os termos foram reunidos para validação utilizando uma base de dados do Stack Overflow para encontrar os resultados sobre accuracy, precision, recall e F_1 -score do novo dicionário. O novo dicionário léxico apresenta 79% de acurácia e precisão, com 78% de Recall e $F1$ -score com um intervalo de polaridade menor do que o dicionário original.*

1. Introdução

A análise de sentimentos é um método de mineração de texto útil para processar conteúdo textual e filtrar os resultados com métodos de análise para obter informações relevantes e significativas [Liu 2015]. Para a manutenção de software, analisar as emoções ou sentimentos dos desenvolvedores através da mineração de textos de bugs/issues ou commits pode se tornar um método capaz de auxiliar a compreensão no gerenciamento do projeto [Murgia et al. 2014].

O processo de classificação textual demanda o uso de ferramentas de suporte, para que seja possível alcançar escala. Dentre as ferramentas existentes, destaca-se a SentiS-

trengh¹, que analisa sentimentos em pequenos trechos de texto [Thelwall et al. 2010]. Com o propósito de analisar o contexto inerente a Engenharia de Software, foi desenvolvida a ferramenta SentiStrength-SE, que faz uso de uma base de dados composta por palavras comuns em Desenvolvimento de Software [Islam and Zibran 2018]. Quando tal conceito é aplicado a ferramentas que desenvolvedores utilizam para troca de textos, é possível relacionar o sentimento do desenvolvedor à sua tarefa.

Neste estudo, analisamos a polaridade de 559 palavras, *emoticons* e expressões idiomáticas contidas no dicionário léxico da ferramenta Sentistrength-SE. Os dados foram obtidos a partir da participação de 48 desenvolvedores de software em um experimento. A alocação dos participantes em grupos considerou os dados do formulário de caracterização, que classificou a experiência em desenvolvimento de software, fluência em inglês e habilidade no tema de análise de sentimentos. A escala de 1 à 5 foi utilizada para descrever os grupos com base nesta caracterização, sendo o *score* 1 atribuído aos participantes com pouca experiência, fluência e habilidade nos grupos apresentados e o *score* 5 aqueles com maior experiência, fluência e habilidade. O estudo foi orientado pelas três perguntas de pesquisa descritas a seguir:

QP1. Os participantes concordam com a classificação das polaridades da ferramenta Sentistrength-SE? O estudo busca mensurar a concordância dos participantes nas rotulações pré-definidas das palavras, *emoticons* e expressões idiomáticas apresentadas no dicionário da ferramenta Sentistrength-SE. Estes termos utilizados na ferramenta foram validados pelos participantes, a partir de uma escala de polaridade variando de -2 a +2. Tal polaridade é comparada com a classificação já apresentada pela ferramenta para encontrar ocorrências de respostas semelhantes. A partir destes resultados, o grau de confiabilidade desta classificação é contabilizado pela concordância do participante com a polaridade rotulada nas respostas. Se houver concordância, então o participante aceita a rotulação do termo de acordo com a classificação, porém rotulações diferentes resultam na discordância da polaridade desta classificação.

QP2. Qual o comportamento dos resultados e a concordância da classificação do dicionário da ferramenta nos grupos de participantes selecionados pela experiência em desenvolvimento de software, fluência no idioma do dicionário e habilidade em análise de sentimentos? O objetivo é encontrar os valores obtidos por cada grupo de participantes que colaboraram na análise do dicionário léxico da ferramenta. É possível que a tendência nas respostas possa trazer algum padrão para elucidar os resultados de concordância e discordância das palavras, *emoticons* e expressões idiomáticas.

QP3. O Dicionário léxico gerado pela classificação dos participantes melhora a acurácia na análise de sentimentos de textos reais? O objetivo do experimento é trazer uma validação dos 559 termos léxicos através de uma nova rotulação destes termos por desenvolvedores de software de todos os graus de experiência. Para responder esta pergunta de pesquisa utilizamos a aferição deste novo dicionário rotulado ao final do processo de coleta do experimento com a análise da base de dados Stack Overflow². As métricas *accuracy*, *precision*, *recall* e *f1-score* do novo dicionário são aferidas.

¹<http://sentistrength.wlv.ac.uk/>

²<http://stackoverflow.com>

2. Metodologia

O estudo foi dividido em três etapas: planejamento, execução e criação/validação do dicionário léxico, conforme descrito a seguir. O estudo piloto obteve respostas de 10 participantes e o experimento contou com a participação de 48 pessoas. Todo o pacote experimental, incluindo os formulários, termos léxicos, tabelas e gráficos suplementares, bem como os dados obtidos, encontra-se disponível em [Menezes et al. 2020].

Etapa 1 - Planejamento: Inicialmente, levantamos um conjunto de 560 termos léxicos com palavras, *emoticons* e expressões idiomáticas do dicionário léxico da ferramenta, mas a palavra 'bg' foi removida por apresentar possível interpretação ambígua na validação e classificação. Assim, o conjunto validado consta de 559 palavras, *emoticons* e expressões idiomáticas. Exemplos de classificação, na frase “*It’s a good feature*”, a palavra “*good*” possui peso 2, então a sentença recebe peso positivo igual a 2 [Islam and Zibran 2018]. Para o *emoticon* “ :) ”, a polaridade é positiva com peso 1. Na expressão idiomática “*Shock horror*” a polaridade é negativa com peso -2.

Para reduzir a granularidade e evitar ambiguidade na rotulação do experimento, os pesos -4 e -3 foram mapeados para o peso -2. Os pesos -2 e -1 foram mapeados para o peso -1. Analogamente, os pesos +1 e +2 foram atribuídos ao peso +1, e os pesos +3 e +4 foram atribuídos ao peso +2. Os termos léxicos foram distribuídos em ordem alfabética em 9 grupos para serem validados e classificados. Um grupo foi utilizado para o experimento piloto e os 8 grupos restantes para o experimento, conforme apresentado na Tabela 1.

Grupo	C		D		E		F		G		H		I		J	
Seção	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2
Palavras	28	27	28	27	28	27	28	27	28	27	28	27	28	27	28	27
Emoticons	3	4	3	4	3	4	3	4	3	3	3	3	3	3	3	3
Exp. Idiom.	4	4	4	4	4	4	4	4	4	5	4	5	4	5	4	5
Participantes	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3

Tabela 1. Dados do Experimento e do Estudo Piloto

Cada grupo de participantes realizou as tarefas seguindo uma dada ordem: (i) **classificação e validação**, ou (ii) **validação e classificação**, conforme detalha a Tabela 2. Na tarefa de **classificação**, o participante rotula o termo léxico, atribuindo-lhe um valor no intervalo de -2 a +2. Na tarefa de **validação**, o termo léxico é rotulado pelo participante que concorda ou não com o rótulo apresentado. Ao discordar, imediatamente é sugerido que o participante apresente uma nova classificação utilizando o intervalo.

Um estudo piloto foi executado, com o objetivo de avaliar os instrumentos do estudo experimental. Para a realização do estudo piloto, dois formulários foram criados. No primeiro formulário, o participante começa a fase de classificação e logo após é realizada a fase de validação. O segundo formulário é realizado de forma inversa, o participante começa a validar e depois classificar novos termos léxicos. Cada formulário foi respondido por cinco pessoas. Ambos os formulários incluem 110 palavras, 20 emoticons e

Tarefas	Grupos
Classificação e Validação	C1, D1, E1, F1, G1, H1, I1, J1
Validação e Classificação	C2, D2, E2, F2, G2, H2, I2, J2

Tabela 2. Ordem das tarefas realizadas no experimento.

18 expressões idiomáticas. O resultado do piloto apresentou 103 respostas iguais aos da ferramenta e 44 respostas diferentes.

Etapla 2 - Execução: O experimento utilizou os formulários descritos no estudo piloto. Um ajuste significativo diz respeito a quantidade de palavras que o participante deveria classificar e validar. A quantidade de palavras, *emoticons* e expressões idiomáticas presentes no piloto foi reduzida pela metade para o experimento, dado que os participantes do piloto relataram fadiga na execução da tarefa com um número elevado de questões. As 440 palavras, 52 emoticons, 68 expressões idiomáticas do dicionário léxico restantes da ferramenta foram divididos nos oito grupos citados anteriormente.

Etapla 3 - Criação e Validação do Novo Léxico: A partir dos dados obtidos no experimento foi possível classificar as polaridades do dicionário da ferramenta, onde a polaridade final de cada palavra, emoticon ou expressão idiomática foi obtida através da maior votação dos participantes. Em casos de empates, considerou-se o valor de maior peso, para ampliar a precisão dos dados. Caso os resultados apresentados tivessem polaridades distintas, optamos por escolher a polaridade encontrada no dicionário da ferramenta previamente extraída. Para validar o léxico obtido do experimento, utilizamos a base de dados *gold standard*³ [Calefato et al. 2018] com 4.423 posts com perguntas, respostas e comentários do *Stack Overflow*. Os dados foram pré-processados, onde os elementos do texto que não continham sentimentos foram descartados, como URLs, trechos de códigos e tags. A performance do novo dicionário léxico foi avaliada utilizando a comparação dos resultados da ferramenta SentiStrength-SE. O SentiStrength-SE Alterado também foi avaliado. Este dicionário consiste exatamente nos termos léxicos do SentiStrength-SE apenas com escala de polaridade modificada de -2 a +2. Como critério de avaliação do novo dicionário léxico foram utilizadas as métricas *accuracy*, *precision*, *recall* e *f₁-score*.

3. Resultados

O experimento coletou 559 respostas de 48 participantes. Os participantes que começaram o estudo na tarefa de **validação** do dicionário levaram em média 7 minutos e 10 segundos, enquanto que os participantes que começaram o estudo na tarefa de **classificação** levaram em média 3 minutos e 58 segundos para a execução do estudo. Os participantes que começaram com a tarefa de **classificação** levaram em média 5 minutos e 15 segundos e 4 minutos e 38 segundos para **validação** do dicionário. As 559 palavras, emoticons e expressões idiomáticas combinadas foram validadas pelos participantes. Tais dados foram analisados na **QP1** e na **QP2**. Em seguida, o dicionário léxico foi avaliado utilizando uma base de dados do Stack Overflow para avaliação da acurácia e discussão da **QP3**.

QP1. Os participantes concordam com a classificação das polaridades da ferramenta SentiStrength-SE? Os participantes que iniciaram com a tarefa de validar a classificação do dicionário da ferramenta concordaram em média com 55,51% das palavras, 59,72% dos emoticons e 71,88% das expressões idiomáticas. Por outro lado, os participantes que iniciaram com a tarefa de classificação concordaram com 56,94% das palavras, 63,54% dos emoticons e 69,79% das expressões idiomáticas. Assim, as expressões idiomáticas obtiveram mais concordância tanto na fase de classificação quanto na fase de validação. Todos os elementos léxicos obtiveram concordância acima de 50%.

³https://github.com/collab-uniba/Senti4SD/blob/master/Senti4SD_GoldStandard_and_DSM/

QP2. Qual o comportamento dos resultados e a concordância da classificação do dicionário da ferramenta nos grupos de participantes selecionados pela experiência em desenvolvimento de software, fluência no idioma do dicionário e habilidade em análise de sentimentos? Neste estudo, os grupos foram separados considerando-se o formulário de caracterização com as tarefas atribuídas aos participantes de classificação e validação. Todos os resultados dos grupos apresentados no estudo estão representados nas Tabelas 3, 4 e 5.

	Tarefa	Grau de experiência				
		1	2	3	4	5
Desenvol. de software	Classificação	43,61%	-	42,83%	36,26%	43,69%
	Validação	74,25%	-	56,31%	58,95%	51,67%
Fluência no idioma	Classificação	37,04%	34%	37,15%	40,37%	47,59%
	Validação	28,57%	32,54%	65,86%	55,39%	56,08%
Experiência no tema	Classificação	44,52%	39,65%	36,74%	37,5%	66,67%
	Validação	54,18%	59,46%	59,65%	42,59%	35,71%

Tabela 3. Concordância na classificação de palavras por domínio e grau.

O grupo que obteve maior concordância com a classificação das palavras apresentadas na ferramenta foi o que realizou a tarefa de validação e possui Experiência de Desenvolvimento de Software grau 1 com 74,25% de acerto. O segundo grupo que alcançou o melhor resultado foi o que realizou a tarefa de classificação e possui maior experiência com análise de sentimentos. O terceiro grupo que melhor realizou a classificação foi o que possuía fluência intermediária no idioma (Grau 3) e realizou a validação. A maior discordância no experimento foi encontrado no grupo de validação que informou menor fluência no idioma (Grau 1), conforme dados apresentados na Tabela 3.

	Tarefa	Grau de experiência				
		1	2	3	4	5
Desenvol. de software	Classificação	30,56%	-	72,22%	40%	40,48%
	Validação	100%	-	67,59%	53,89%	59,13%
Fluência no idioma	Classificação	25%	88,89%	46,53%	35,18%	50,60%
	Validação	0%	50%	56,25%	68,98%	63,69%
Experiência no tema	Classificação	53,07%	37,5%	41,67%	50%	50%
	Validação	71,05%	50,69%	63,10%	25%	66,67%

Tabela 4. Concordância nos emoticons por domínio e grau.

Na análise dos emoticons, os três grupos que obtiveram maiores concordâncias com a classificação da ferramenta foram, respectivamente, os grupos que possuíram experiência de desenvolvimento de software grau 1 na tarefa de validação com 100%, fluência no idioma grau 02 na tarefa de classificação com 88,89% e experiência em desenvolvimento de software grau 03 na tarefa de classificação com 72,22%. A maior discordância com a classificação foi encontrada nos grupos de fluência de idioma na validação grau 1 com 0% e classificação de desenvolvimento de software grau 1 com 30,56%.

Ao avaliar as expressões idiomáticas, o grupo com tarefa de validação com fluência no idioma grau 1 obteve maior concordância com a classificação, seguido do grupo com tarefa de validação com experiência em desenvolvimento de software grau 1 com 91,67% de concordância. Os grupos que discordaram das classificações das expressões idiomáticas foram os que possuíam maior experiência em análise de sentimento e realizaram a tarefa de validação, com 12,50%, e classificação, com 25%.

QP3. O Dicionário léxico gerado pela classificação dos participantes melhora a acurácia na análise de sentimentos de textos reais? O dicionário léxico com as polaridades

	Tarefa	Grau de experiência				
		1	2	3	4	5
Desenvol. de software	Classificação	58,33%	-	50%	50%	45,24%
	Validação	91,67%	-	77,78%	75%	61,9%
Fluência no idioma	Classificação	50%	33,33%	43,75%	44,44%	60,71%
	Validação	100%	41,67%	85,42%	66,67%	67,86%
Experiência no tema	Classificação	46,05%	47,92%	55,36%	25%	50%
	Validação	67,11%	66,67%	89,29%	12,50%	50%

Tabela 5. Concordância nas expressões idiomáticas por domínio e grau.

obtidas foi submetido à validação utilizando a base de dados de textos do Stack Overflow³. Cada palavra, emoticon e expressão idiomática foi considerada de acordo com a polaridade mais votada. É importante ressaltar que, nos casos em que os resultados apresentados obtiveram polaridades distintas das selecionadas nos formulários, houve a escolha da polaridade encontrada na base da ferramenta previamente extraída. O novo dicionário léxico apresenta 79% de Acurácia e Precisão, com 78% de *Recall* e *F₁-score*. Isto representa uma acurácia e precisão menor que os resultados apresentados pelo léxico do *SentiStrength-ES*, que utiliza uma escala maior de polaridade do que a utilizada no Novo Dicionário. Para o *SentiStrength-ES* Alterado houve resultados melhores como o aumento de 6% da acurácia, 3% da precisão na classificação, 7% no *recall* e *F₁-score*, conforme apresentado na Tabela 6.

Léxico	Accuracy	Precision	Recall	F ₁ -score
SentiStrength-ES	81%	81%	82%	81%
SentiStrength-ES Alterado	73%	76 %	71 %	71%
Novo Dicionário	79%	79%	78%	78%

Tabela 6. Resultados obtidos na validação do dicionário léxico

Os valores das métricas ficaram próximo dos originais, variando entre 2% e 4%, o que pode ser considerado valores próximos. Em relação com o dicionário da *SentiStrength* Alterado houve um aumento significativo de acurácia, precisão, *recall* e *F₁-score*. A validação e classificação dos termos trouxe um maior ganho para o Novo Dicionário, sendo o SentiStrength-ES Alterado operar na mesma escala. A diferença da nova avaliação das palavras, emoticons e expressões idiomáticas foi positiva para ganho na acurácia, precisão e outros indicadores, conforme apresentado na Tabela 6.

4. Ameaças à validade

Validade Interna: No estudo piloto, os participantes relataram exaustão na execução das tarefas. Como efeito, houve a redução da quantidade de questões, de 147 para 70. Houve também a redução das escalas de classificação dos sentimentos, que era de -4 até +4 que passou para -2 à +2. O método de coleta de resultados envolveu a utilização de um questionário *online* compartilhado com os participantes. Este recurso não oferece condições para o monitoramento do ambiente, nem para conduzir que todas as respostas sejam realizadas no mesmo dia e local. Portanto, 16 modelos diferentes de formulários com a mesma estrutura foram aplicados aos participantes.

Validade Externa: A área de atuação que o estudo busca alcançar é a Engenharia de Software, portanto a condução do estudo e a construção e validação do dicionário léxico foi projetado para a área em questão. A generalização dos resultados e a utilização do

dicionário léxico em outros contextos e áreas não são garantias de resultados semelhantes aos apresentados neste estudo.

Validade de Construção: Emoticons que não são populares no dia a dia do participante pode levar a uma interpretação equivocada da sua polaridade. O estudo considerou 72 emoticons, sendo 20 emoticons no piloto e 52 no experimento. Deste total apenas 1 emoticon não foi corretamente identificado. Para números do estudo temos que ele significa 1,42% do total da base que utilizamos, portanto a classificação deste emoticon por critério de desempate foi considerada a classificação da ferramenta. De modo análogo, é possível que tal equívoco ocorra na classificação das palavras, visto que o participante pode não reconhecer os termos e então classificar de maneira errada, assim prejudicando a classificação do texto. A fluência no idioma é importante em todo o processo.

5. Trabalhos Relacionados

Os autores da SentiStrength-SE [Islam and Zibran 2018] realizaram um experimento com três alunos de graduação em Ciência da Computação com pelo menos três anos de experiência em desenvolvimento de software. Na primeira etapa, os autores removeram 716 palavras do dicionário léxico da SentiStrength [Thelwall et al. 2010] que representam termos técnicos no domínio de Engenharia de Software e não expressam sentimento no contexto desse domínio. O restante do dicionário foi avaliado pelos participantes. Em seguida, as palavras que foram classificadas como neutras foram excluídas do dicionário.

O trabalho de [Calefato et al. 2018] buscou, através de um experimento controlado, criar uma base de dados com questões, respostas e comentários do *Stack Overflow* rotulada com as polaridades Positiva, Neutra e Negativa. O estudo contemplou uma atividade de treinamento com os participantes. Cada um deles deveriam rotular 25 posts em 30 minutos. Os posts foram rotulados em positivo, negativo, neutro e misturado. Na atividade do experimento, os participantes foram organizados em quatro grupos de três participantes cada, e cada participante classificou 100 *posts*. No experimento, cada participante classificou 500 *posts* e mais uma vez, os posts foram rotulados por três codificadores, gerando 2000 novos posts rotulados.

O presente estudo utilizou uma maior quantidade de termos léxicos para validação a partir de um experimento utilizando formulários online com uma maior quantidade de participantes. No total 48 pessoas participaram do experimento para validar 559 palavras, *emoticons* e expressões. Essa heterogeneidade traz maior possibilidade de respostas na validação do dicionário.

6. Considerações Finais

O presente estudo validou as polaridades e pesos do dicionário léxico da ferramenta SentiStrength-SE. A validação foi realizada através de um experimento com 48 participantes, que possuem diferentes níveis de experiência em desenvolvimento de software, fluência no idioma e conhecimento no tema de análise de sentimentos.

A análise dos artefatos do repositório é importante para a evolução e manutenibilidade do projeto. Em [Boechat et al. 2019], a análise utilizou a classificação das polaridades das *issues* reabertas com apoio fundamental da ferramenta SentiStrength para encontrar respostas sobre a previsão de reabertura das *issues* fechadas. Com um dicionário validado e ajustado, a análise de sentimentos dos dados pode trazer resultados mais

precisos. O dicionário léxico de Engenharia de Software já existe, mas a construção em cima da validação de termos por desenvolvedores de software é fundamental para melhorar os resultados de classificação da ferramenta. Assim, os termos léxicos validados e construídos neste dicionário podem ajudar na classificação de sentimentos tanto de *issues*, como de outros artefatos relacionados com a manutenibilidade e evolução do repositório de *software*. A validação utilizando a base de dados do Stack Overflow, que é uma base relacionada com desenvolvimento de software, foi realizada neste estudo com 79% de acurácia, 78% de *recall*, 79% de precisão, e 78% de *f1 - score*. Com os números apresentados utilizando a base do Stack Overflow, o dicionário léxico deste estudo pode ser utilizado como base para analisar textos voltados para desenvolvimento de software, devido a validação ser realizada diretamente por desenvolvedores que possuem experiências diversas na área. Assim, este dicionário auxilia na análise de outros elementos importantes para a evolução e manutenção de repositórios, como as *issues*, *pull-request*, *commits* e outros artefatos.

Para trabalhos futuros é esperado a expansão deste dicionário léxico com a validação de especialistas na área de análise de sentimentos. Uma outra direção a seguir é a classificação de diferentes bases de dados, contemplando um número maior de textos de *issues* e *commits*.

Agradecimentos: O presente trabalho foi realizado com apoio da FAPESB TO JCB0060/2016, TO BOL2743/2016, e da CAPES – Cód. de Financiamento 001 (Nº 88887.494000/2020-00).

Referências

- Boechat, G., Júnior, J. M., Machado, I., and Mendonça, M. (2019). Análise de sentimentos em discussões de issues reabertas do github. In *Anais do VII Workshop on Software Visualization, Evolution and Maintenance (VEM)*, pages 13–20. SBC.
- Calefato, F., Lanubile, F., Maiorano, F., and Novielli, N. (2018). Sentiment Polarity Detection for Software Development. In *Proceedings of the 40th International Conference on Software Engineering, ICSE*, pages 128–128, NY, USA. ACM.
- Islam, M. R. and Zibran, M. F. (2018). SentiStrength-SE: Exploiting domain specificity for improved sentiment analysis in software engineering text. *J. of Systems and Software*, 145:125 – 146.
- Liu, B. (2015). *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*. C.U.P.
- Menezes, H., Boechat, G., Júnior, J. M., and Machado, I. (2020). Validação e construção de um dicionário léxico para auxiliar a análise de sentimentos em repositórios de projetos de software (material suplementar). Zenodo
<http://doi.org/10.5281/zenodo.4029777>.
- Murgia, A., Tourani, P., Adams, B., and Ortu, M. (2014). Do Developers Feel Emotions? An Exploratory Analysis of Emotions in Software Artifacts. In *Proceedings of the 11th Conf. on Mining Software Repositories(MSR)*, page 262–271, NY, USA. ACM.
- Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., and Kappas, A. (2010). Sentiment Strength Detection in Short Informal Text. *J. Am. Soc. Inf. Sci. Technol.*, 61(12):2544–2558.