Uma avaliação de ferramentas de análise de sentimentos aplicadas a comentários da plataforma GitHub

Giuseppe Portolese[1], **Guilherme A. M. da Cruz**[1],
Elisa H. M. Huzita[1], **Valéria D. Feltrim**[1]

[1]Departamento de Informática – Universidade Estadual de Maringá (UEM)
Av. Colombo, 5.790 – 87020-900 – Maringá - PR - Brasil

{giuportolese, guilherme.maldonado.cruz}@gmail.com
{emhuzita, vfeltrim}@din.uem.br

Abstract. *Distributed software development has become frequent and the interaction between those involved, which is often influenced by social and cultural aspects, reflects in the performance of the teams. Sentiment analysis has been used to capture subjective information and get a better understanding of the interactions of these teams. Therefore, it is interesting to evaluate the performance of available tools when applied to that domain. In this work nine sentiment analysis tools were evaluated using GitHub comments manually annotated according to their polarity. Results showed that SentiStrength performed best among the evaluated tools, but with average performance below 50%.*

Resumo. *O desenvolvimento distribuído de software tem se tornado frequente e a interação entre os envolvidos, muitas vezes influenciada por aspectos sociais e culturais, reflete no desempenho das equipes. A análise de sentimentos vem sendo empregada para capturar informações subjetivas e obter um maior entendimento das interações dessas equipes. Portanto, é interessante avaliar o desempenho das ferramentas disponíveis quando aplicadas a esse domínio. Neste trabalho nove ferramentas de análise de sentimentos foram avaliadas usando comentários extraídos da plataforma GitHub e que foram manualmente anotados quanto à polaridade. Os resultados mostraram que a ferramenta SentiStrength se saiu melhor, porém com desempenho médio abaixo de 50%.*

1. Introdução

A distribuição geográfica dos membros caracteriza o desenvolvimento distribuído de software (DDS) e tem como objetivo trazer benefícios, como a facilidade em encontrar mão de obra qualificada, redução dos custos e agilidade na entrega dos produtos por meio da utilização do desenvolvimento *follow-the-sun* [O'Conchuir et al. 2006]. Contudo, a distância acrescenta desafios ao desenvolvimento de software que podem ser divididos, de forma geral, em socioculturais e técnicos. Desafios socioculturais englobam comunicação limitada e diferenças culturais, entre outros. Os desafios técnicos, por sua vez, referem-se a problemas com a rede, segurança das informações, processos/ferramentas de trabalhos diferentes, entre outros [Herbsleb and Moitra 2001, Sengupta et al. 2006].

Devido aos desafios oriundos da distribuição geográfica, diversos estudos e ferramentas visam auxiliar essas equipes e diminuir os efeitos da distância, tais como fóruns,

wikis, ferramentas de mensagens instantâneas, de videoconferência, para acompanhamento de atividades, monitoramento da equipe e de versionamento. Alguns desses estudos têm usado a análise de sentimentos sobre artefatos e comunicações de equipes desenvolvimento como forma de capturar informações subjetivas e, assim, buscar um maior entendimento a respeito das interações e emoções dos membros dessas equipes. Por exemplo, Guzman et al. [2014] empregam análise de sentimentos em artefatos produzidos ao longo de projetos a fim de obter o clima emocional dos membros envolvidos. Sinha et al. [2016] analisaram *commit logs* do GitHub usando análise de sentimentos e buscaram relações com os dias da semana e a quantidade de modificações. Cruz et al. [2016] usaram análise de sentimentos como parte de um *framework* para estimar a confiança entre membros de equipes de DDS.

Embora ferramentas de análise de sentimentos estejam sendo empregadas nesses e em outros estudos no contexto do desenvolvimento distribuído de software, em geral, essas ferramentas não foram criadas visando a aplicação nesse domínio. O mais comum é que as ferramentas tenham foco na análise de textos provenientes de redes sociais e *reviews*, e, por isso, tendem a não ter o mesmo desempenho quando aplicadas ao domínio de DDS [Tourani et al. 2014, Sinha et al. 2016]. Jongeling et al. [2015] argumentam ainda que dependendo da escolha da ferramenta, o estudo pode obter conclusões contraditórias, uma vez que além da baixa precisão, as ferramentas podem discordar entre si.

Tendo em vista esse cenário, este trabalho teve por objetivo avaliar diferentes ferramentas de análise de sentimentos no contexto das comunicações entre membros de equipes distribuídas. Como fonte de dados utilizamos a plataforma GitHub, mais especificamente, comentários feitos em *pull requests*, uma vez que esse tipo de comentário tem sido utilizado em vários trabalhos no contexto de DDS [Sinha et al. 2016, Guzman et al. 2014, Cruz et al. 2016]. O restante deste artigo está organizado em quatro seções. Na Seção 2 são descritos trabalhos que empregaram análise de sentimentos no contexto de desenvovimento de software. Na Seção 3 é descrita a metodologia, incluindo os dados utilizados e as ferramentas avaliadas. Na Seção 4 são apresentados os resultados e, por fim, na Seção 5, são apresentadas as conclusões e direções para possíveis trabalhos futuros.

2. Motivação

A análise de sentimentos (AS) é uma área de pesquisa ampla e interdisciplinar que diz respeito ao estudo de opiniões, sentimentos, atitudes e emoções. Dentre as diferentes tarefas tratadas por ferramentas de AS, a mais comum é a de determinar a polaridade de um texto – se positiva, negativa ou neutra – e, em geral, se estabelece em um de três níveis: (i) nível de documento, (ii) nível da sentença e (iii) nível da entidade ou aspectos. A área tem recebido muita atenção por parte da comunidade científica e tem encontrado aplicações em quase todos os domínios [Liu 2012].

No contexto do desenvolvimento de software, pesquisadores vêm empregando análise de sentimentos com o objetivo de entender e capturar aspectos subjetivos relativos à comunicação e ao relacionamento dos membros dessas equipes. Essa análise é especialmente interessante no caso das equipes de DDS, uma vez que a comunicação mediada por computador, que é característica dessas equipes, é uma fonte abundante de dados para a análise de sentimentos.

Guzman et al. [2014] usaram análise de sentimentos em comentários de *com-*

mits de 90 projetos da plataforma GitHub, desenvolvidos em diferentes linguagens de programação. Os autores utilizaram a ferramenta SentiStrength [Thelwall 2013] para classificar a polaridade dos comentários e analisaram a existência de relação entre as médias de polaridade observadas e a linguagem de programação usada no projeto, o dia da semana em que os *commits* foram criados, a distribuição geográfica, e a aprovação do projeto. Os resultados mostraram que projetos em Java tiveram a polaridade média levemente negativa em comparação a outras linguagens e que *commits* criados nas segundas-feiras tendem a ser mais negativos. Não foi encontrada relação entre a distribuição geográfica e a polaridade dos comentários, mas notou-se que, quanto maior é a distribuição, maior é a força da polaridade nos comentários positivos. Também não foi encontrada relação entre a aprovação do projeto e a polaridade dos comentários, mas foi observada uma correlação positiva fraca entre a média dos comentários positivos e a aprovação do projeto.

Tourani et al. [2014] analisaram a presença e a evolução de sentimentos negativos e positivos nas listas de e-mails de desenvolvedores e usuários dos dois maiores projetos da Apache: Tomcat e Ant. Um subconjunto desses e-mails foi anotado manualmente em relação à polaridade e essa anotação foi então comparada à classificação fornecida pela ferramenta SentiStrength. Os autores relatam que a ferramenta obteve baixa precisão quando comparada à anotação manual e destacam a necessidade de costumização das ferramentas de AS ao domínio de desenvolvimento de software. Os resultados da análise mostraram que a polaridade dos comentários evolui com o tempo, havendo momentos de picos positivos e negativos. Além disso, constatou-se que usuários e desenvolvedores apresentam sentimentos diferentes durante diferentes fases do projeto.

Sinha et al. [2016] analisaram *commit logs* do GitHub que foram disponibilizados como parte do MSR 2016 *challenge*. Assim como nos trabalhos anteriores, a SentiStrength foi usada para determinar os valores de polaridade. Os resultados mostraram 74,74% comentários neutros, 7,19% positivos e 18,05% negativos e que, em projetos maiores, a diferença entre a quantidade de comentários positivos e negativos é maior do que em projetos menores. Em termos das polaridades para os dias da semana, as terças-feiras tiveram comentários com polaridade negativa mais alta. Além disso, foi encontrada forte correlação entre a quantidade de arquivos alterados e o sentimento desses *commits*.

Cruz et al. [2016] utilizaram análise de sentimentos como parte de um *framework* automático para estimar a confiança entre membros de equipes DDS. A estimativa é feita por meio da extração de indícios de confiança que podem ser observados nas interações de um sistema de versionamento. Entre os indícios considerados estão: confiabilidade, tom positivo da comunicação, mímica de vocabulário, aceitação de conhecimento, colaboração e delegação. A extração dos indícios é feita a partir dos valores fornecidos pela ferramenta SentiStrength aplicada em comentários de *pull requests* nas quais os membros interagem, e de outras informações extraídas do GitHub.

Jongeling et al. [2015] avaliaram o desempenho das ferramentas SentiStrength, NLTK, Alchemy e *Stanford NLP sentiment analyser*, quando aplicadas a comentários do repositório da *Apache software foundation*. Os 392 comentários usados haviam sido anotados manualmente como parte do trabalho de Murgia et al. [2014]. As ferramentas NLTK e SentiStrength obtiveram os melhores resultados na avaliação, embora abaixo dos reportados para outros domínios. Assim como Jongeling et al. [2015], este trabalho também se propõe a avaliar ferramentas de análise de sentimentos aplicadas a comentários

no domínio de desenvolvimento de software, porém usando comentários extraídos da plataforma GitHub e ampliando a quantidade de ferramentas avaliadas para nove.

3. Avaliação das Ferramentas de Análise de Sentimentos

3.1. Dados

Uma vez que o interesse deste trabalho está na análise de sentimentos aplicada à comunicação entre membros de equipes distribuídas, optamos por utilizar como fonte de dados comentários feitos em *pull requests* de projetos hospedados na plataforma GitHub. Assim, foram extraídos automaticamente 350 comentários a partir de quatro projetos, escolhidos por apresentarem um número elevado de comentários. Cada comentário foi manualmente pré-processado para separação das sentenças, uma vez que as ferramentas avaliadas fazem classificação sentencial.

Após o pré-processamento, as 2.041 sentenças resultantes foram manualmente classificadas como positivas, negativas ou neutras, de acordo com a polaridade da emoção expressa na sentença avaliada. A classificação manual de todas as sentenças foi feita por um anotador com formação em Ciência da Computação. O número e o percentual de sentenças classificadas para cada valor de polaridade são mostrados na Tabela 3.1.

Uma característica particular de comentários como os que foram extraídos do GitHub é a presença de trechos de código-fonte, bem como observações sobre o seu funcionamento, como parte do texto do comentário. Nesses casos, o anotador procedeu a classificação da seguinte forma: sentenças que se referiam ao correto funcionamento do código foram classificadas como positivas; sentenças em que usuário reportava o mal funcionamento do código foram classificadas como negativas; e sentenças que correspondiam apenas a trechos de código-fonte foram classificadas como neutras.

Classificação	**#Sentenças**	**%Sentenças**
Positiva	260	12,7
Neutra	1657	81,2
Negativa	124	6,1
Total	**2.041**	**100**

Tabela 1. Quantidade de sentenças por polaridade

Conforme mostra a Tabela 3.1, 81% das sentenças analisadas foram classificadas como neutras, enquanto 19% foram consideradas positivas ou negativas, sendo que a proporção de sentenças positivas e negativas é de aproximadamente dois para um. Essa distribuição é similar as observadas por Jongeling et al. [2015] e Murgia et al. [2014] e se deve ao fato de muitos dos comentários descreverem aspectos técnicos do desenvolvimento e não expressarem emoção de forma perceptível ao anotador. Há também comentários em que a polaridade (positiva ou negativa) pode ser identificada com maior facilidade, devido ao uso de expressões características de uma polaridade, emoticons e formas de escrita que procuram imitar a linguagem oral, como reticências, pontos de exclamação e onomatopéias, no entanto, esse não é o caso mais frequente.

Para avaliar a reprodutibilidade da anotação manual, 215 comentários foram anotados por um segundo anotador, também com formação em Ciência da Computação. A concordância entre os dois anotadores, estimada por meio da estatística *Kappa*

[Cohen 1960], foi de $0,46$, evidenciando a subjetividade da classificação de polaridade. Uma maior concordância foi observada para as sentenças positivas ($K = 0,59$) e negativas ($K = 0,45$), o que mostra uma maior dificuldade em distinguir entre sentenças positivas/negativas e neutras do que entre sentenças positivas e negativas.

3.2. Ferramentas

As seguintes ferramentas de análise de sentimentos foram avaliadas neste estudo:

- **Análise por *emoticons*** [Park et al. 2013]: Utiliza uma tabela que relaciona os *emoticons* mais populares com a polaridade a qual são atribuídos com maior frequência. A tabela empregada foi a de Gonçalves et al. [2013] e considerou-se a polaridade da sentença como sendo a mesma do primeiro *emoticon* encontrado.
- **SentiStrength** [Thelwall 2013]: Analisa o texto de entrada e atribui a cada sentença pontuações referentes às polaridades positiva e negativa, usando um modelo baseado em aprendizado de máquina. Embora tenha sido desenvolvida para a análise de textos curtos, como os do *Tweeter*, tem sido empregada em vários trabalhos no contexto do desenvolvimento de software [Cruz et al. 2016, Jongeling et al. 2015, Guzman et al. 2014]. Neste trabalho a ferramenta foi configurada para retornar os valores positivo, negativo ou neutro a cada sentença.
- **SentiWordNet** [Baccianella et al. 2010]: É uma base de dados lexical para a mineração de opinião baseada na WordNet , na qual cada *synset* possui valores referentes à objetividade, positividade e negatividade. A base retorna valores entre 1 e -1 para os termos dependendo dos papéis que exercem na frase. Para atribuir polaridade a uma sentença, foi feita a soma dos valores médios retornados para cada termo. Caso a soma fosse igual a 0, atribuiu-se polaridade neutra à sentença; caso contrário, a polaridade foi atribuída com base no sinal do valor da soma.
- **SenticNet** [Cambria et al. 2010]: é uma base de conhecimento de senso comum que usa diferentes técnicas de inteligência artificial para inferir a polaridade de um texto a partir de conhecimento semântico. Dessa forma, a análise se dá em nível conceitual e não apenas em nível léxico e sintático. A base retorna valores de -1 a 1 referente à polaridade do termo pesquisado. A atribuição de polaridade a uma senteça foi feita por meio da média aritimética simples dos termos da frase. Caso o resultado fosse 0, atribuiu-se polaridade neutra; caso contrário, a polaridade foi atribuida levando em consideração o sinal da média obtida.
- **Happiness Index** [Dodds and Danforth 2010]: Analisa a frequência de termos pré-classificados do dicionário ANEW *(Affective Norms for English Words)*, os quais tiveram valores atribuídos por juízes humanos usando uma escala contínua de ”felicidade”. Como a ferramenta atribui valores de 0 a 9 aos termos analisados, para avaliar a polaridade de uma sentença, foi feita a média aritimética simples dos valores obtidos para cada palavra. Caso o valor resultante fosse maior ou igual a 6, atribuiu-se polaridade positiva; caso fosse menor ou igual a 4, atribuiu-se polaridade negativa; no restante dos casos, a polaridade foi considerada neutra.
- **PANAS-t** [Goncalves et al. 2012]: Escala psicométrica para a detecção de humor na plataforma Twitter que aplica uma versão adaptada do *Positive Affect Negative Affect Scale* (PANAS) [Watson et al. 1988] utilizada na psicologia. Para atribuir polaridade a uma sentença, busca-se identifcar se a sentença fala sobre o estado emocional de seu autor; caso contrário, a sentença é considerada neutra. Nos casos

em que a sentença fala sobre um estado emocional, atribui-se a polaridade positiva ou negativa com base no estado emocional identificado.

- **AFINN** [Nielsen 2011] e **Sentiment140 Lexicon** [Mohammad et al. 2013]: Ambas são bases léxicas que possuem valores de polaridade pré-determinados atribuídos a termos da língua inglesa. Para avaliar a polaridade de uma sentença, neste trabalho, foi feita a soma dos valores de polaridade das palavras da sentença; caso o resultado fosse 0, atribuiu-se polaridade neutra; caso contrário, a polaridade foi atribuida levando em consideração o sinal do valor obtido.

Além das ferramentas citadas, também foi avaliado um método combinado, que estima a polaridade da sentença por meio da combinação linear das saídas das oito ferramentas analisadas. Esse método é similar ao proposto por Araujo et al. [2014] para a classificação de *tweets*, mas se diferencia na forma do sistema de pesos e na busca pelos pesos mais apropriados. No método combinado usado neste trabalho, cada ferramenta recebeu três pesos, um para cada polaridade. Dessa forma, dependendo da polaridade aferida pela ferramenta, o peso correspondente foi utilizado. Esses pesos foram estimados por um algoritmo genético, que buscou maximizar o desempenho da classificação em termos da medida-F obtida para cada polaridade.

3.3. Medidas

As feramentas foram avaliadas usando as sentenças resultantes do pré-processamento, conforme descrito na Subseção 3.1, e tiveram o seu desempenho registrado em termos das seguintes medidas, estimadas para cada valor de polaridade p:

- Precisão: total de setenças corretamente classificadas como p sobre o total de sentenças classificadas como p;
- Cobertura: total de setenças corretamente classificadas como p sobre o total de sentenças com polaridade p no conjunto;
- Medida-F: média harmônica dos valores de precisão e cobertura obtidos para p.

Além das medidas por polaridade, também foi calculada a macro-F para cada ferramenta, que corresponde à média aritmética dos respectivos valores de medida-F.

4. Resultados Obtidos

Os resultados obtidos pelas ferramentas avaliadas em termos da macro-F e da precisão, cobertura e medida-F para os três valores de polaridade – positiva, negativa e neutra –, são apresentados na Tabela 2. Como pode ser observado, a ferramenta que obteve o melhor resultado em termos de macro-F foi a SentiStrength (47,8%), seguida pelo método combinado (47%). A ferramenta SenticNet obteve a menor macro-F registrada nas avaliações (18,5%), ficando abaixo de métodos mais simples, como a análise por *emoticons* e *hapiness index*. A baixa macro-F obtida pela SenticNet pode ser explicada analisando-se as distribuição das classificações nas polaridades positiva, negativa e neutra geradas pela ferramenta, que foi de 62,3%, 21,9% e 15,9%, respectivamente, enquanto a distribuição observada na anotação manual foi de 12,7%, 6,1% e 81,2%. Com exceção da SenticNet, as classificações de todas as ferramentas apresentaram distribuição similar a da anotação manual, sendo as sentenças neutras majoritárias e as sentenças negativas minoritárias.

Conforme mostra a Tabela 2, os resultados obtidos com o método combinado ficaram consistentemente entre os melhores quando considera-se cada ferramenta individualmente, porém nunca estão em primeiro lugar em cada ponto individual. Assim, é

possível inferir que o método combinado incorpora os pontos fortes das ferramentas que o compõe, mas também os erros de classificação que essas ferramentas apresentam.

	Precisão			Cobertura			F1			
	Positivo	**Neutro**	**Negativo**	**Positivo**	**Neutro**	**Negativo**	**Positivo**	**Neutro**	**Negativo**	**Macro-F**
SentiWordNet	13,68%	77,45%	6,86%	35,77%	37,72%	30,65%	19,79%	50,73%	11,21%	27,24%
Análise de Emoticons	94,59%	82,76%	100,00%	13,46%	99,94%	2,42%	23,57%	90,54%	4,72%	39,61%
SentiStrength	36,29%	90,39%	15,12%	67,69%	60,71%	54,03%	47,25%	72,64%	23,63%	**47,84%**
SenticNet	12,04%	67,59%	8,52%	58,85%	13,22%	30,65%	19,99%	22,11%	13,33%	18,48%
Happiness Index	15,08%	81,44%	4,65%	23,08%	76,52%	3,23%	18,24%	78,90%	3,81%	33,65%
PANAS-t	25,00%	81,21%	14,29%	1,15%	99,09%	0,81%	2,21%	89,26%	1,53%	31,00%
AFINN	21,65%	84,64%	16,39%	50,38%	61,19%	31,45%	30,29%	71,03%	21,55%	40,96%
Sentiment140	41,53%	83,15%	15,38%	18,85%	93,24%	8,06%	25,93%	87,91%	10,58%	41,47%
Genético	50,00%	84,06%	24,64%	26,92%	92,94%	13,71%	35,00%	88,28%	17,62%	46,96%

Tabela 2. Resumo dos resultados obtidos pelas ferramentas avaliadas

Os gráficos das figuras 1, 2 e 3 mostram a comparação dos valores de precisão, cobertura e medida-F, respectivamente, obtidos pelas ferramentas avaliadas. No gráfico da Figura 1 percebe-se que a análise por *emoticons* teve precisão superior, especialmente para as polaridades positiva e negativa, para as quais a precisão das outras ferramentas é relativamente baixa. No entanto, a cobertura desse método para essas polaridades está entre as mais baixas observadas, conforme mostra a Figura 2. O método combinado teve uma precisão boa comparada às outras ferramentas, mas manteve a mesma tendência de baixa precisão para as polaridades negativa e positiva.

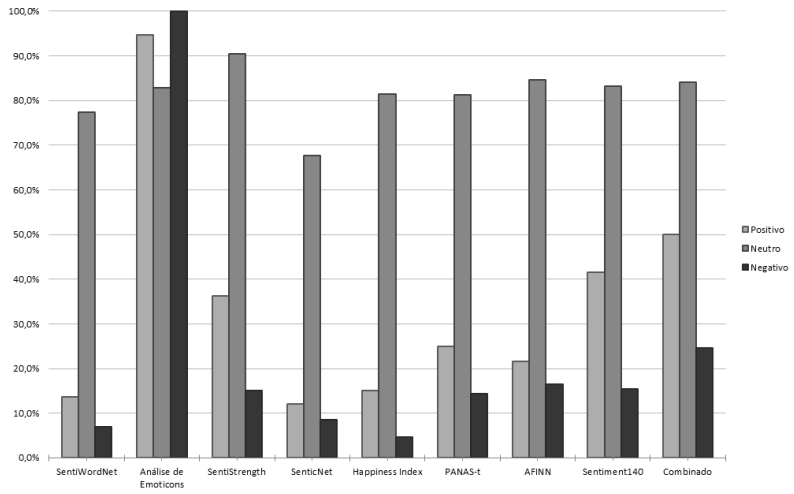

Figura 1. Comparação das ferramentas em termos da precisão

Com relação à cobertura, houve uma discrepância maior no comportamento das ferramentas. Conforme mostra o gráfico da Figura 2, ferramentas com alta cobertura para sentenças neutras tendem a possuir baixa cobertura para as outras duas polaridades (p.e., análise por *emoticons* e PANAS-t), enquanto outras ferramentas apresentaram cobertura mais baixa para sentenças neutras do que para sentenças positivas/negativas (p.e., SentiStrength e SenticNet). Analisando-se a média dos valores de cobertura para as três polaridades, o melhor resultado foi obtido pela SentiStrength ($60,8\% \pm 6,8\%$), seguida pela ferramenta AFINN ($47,7\% \pm 15,1\%$) e pelo método combinado ($44,5\% \pm 42,4\%$). O pior resultado foi obtido pela PANAS-t ($33,7\% \pm 56,6\%$).

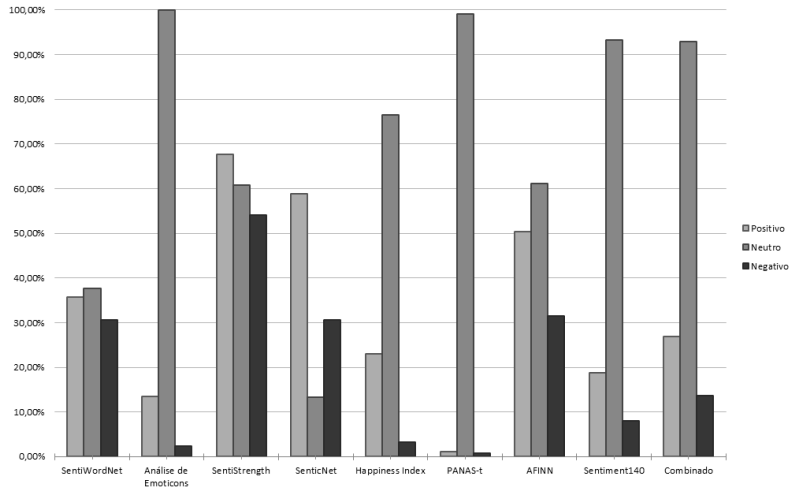

Figura 2. Comparação das ferramentas em termos da cobertura

O gráfico da Figura 3 mostra um comparativo das ferramentas em termos da medida-F. Como pode ser observado, a ferramenta SentiStrength obteve a melhor medida-F para as polaridades positiva e negativa, enquanto a PANAS-t obteve a pior. O método combinado, por sua vez, superou a SentiStrength para sentenças neutras, ficou em segundo lugar para sentenças positivas e em terceiro lugar para sentenças negativas, sendo superado, nesse caso, pela ferramenta AFINN.

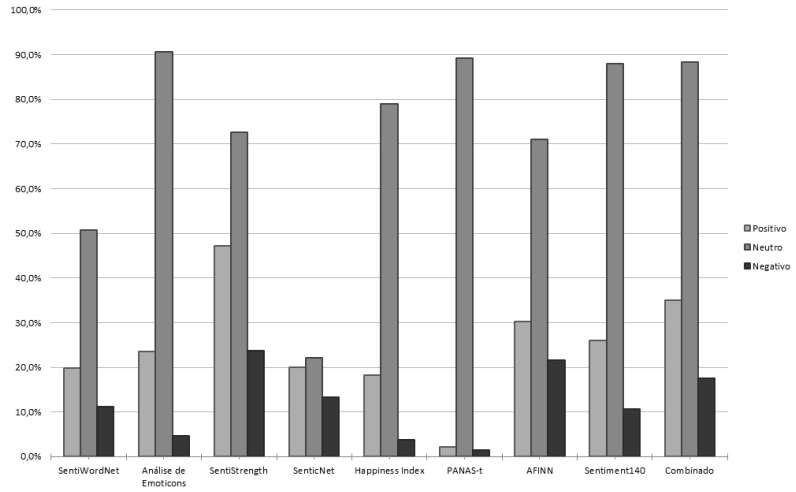

Figura 3. Comparação das ferramentas em termos da medida-F

5. Conclusões e Trabalhos Futuros

O desenvolvimento de software é caracterizado pela forte interação que se estabelece entre as pessoas envolvidas. Essas pessoas podem carregar características oriundas de uma herança socio-cultural, que nem sempre são expressas verbalmente, mas que podem aparecer embutidas na comunicação escrita. Nesse sentido, a análise de sentimentos se mostra como uma ferramenta importante para a captura automática de aspectos subjetivos relacionados à interação e ao relacionamento dos membros das equipes de desenvolvimento. Para isso, é necessário que se conheça o desempenho das ferramentas de análise

de sentimentos disponíveis quando aplicadas a esse domínio, uma vez que a classificação produzida influencia diretamente nas conclusões dos estudos conduzidos com base nessas informações. Cabe ressaltar que a correta identificação e uso de informações acerca da interação e relacionamento entre membros da equipe pode impactar as atividades de um processo de desenvolvimento e, consequentemente, na qualidade do produto final.

Neste trabalho foram avaliadas nove ferramentas de análise de sentimentos aplicadas a comentários da plataforma GitHub, incluíndo ferramentas populares, como a SentiStrength, e um método combinado similar ao proposto por Gonçalves et al. [2014]. A análise dos resultados mostrou que a SentiStrength saiu-se melhor, seguida pelo método combinado, quando se considerou o desempenho médio para os três valores de polaridade considerados. Também foi possível observar que métodos simples como a análise por *emoticons* podem obter valores altos de precisão, embora tendam a ter baixa cobertura. Os resultados desse estudo mostraram uma variação considerável de desempenho entre as ferramentas e que, mesmo para a ferramenta melhor avaliada, o desempenho médio ficou abaixo de 50%. Isso mostra o quanto a escolha da ferramenta de análise de sentimentos pode influenciar os estudos que as utilizam. Além disso, evidencia a necessidade de investigação na área e do desenvolvimento de ferramentas que sejam capazes de capturar as especificidades de textos como os produzidos por equipes de DDS.

Como trabalhos futuros pretende-se melhorar a base de comentários anotados, aumentando o volume de sentenças anotadas, refinando os critérios adotados na anotação das polaridades e realizando experimentos de anotação com mais anotadores. Também pretende-se criar outras versões do método combinado por meio do uso de outros algoritmos para a estimativa de pesos e outras formas de combinar as ferramentas.

Referências

Araújo, M., Gonçalves, P., Cha, M., and Benevenuto, F. (2014). ifeel: A system that compares and combines sentiment analysis methods. In *Proc. of the 23rd Int. Conf. on World wide web Companion*, pages 75–78.

Baccianella, S., Esuli, A., and Sebastiani, F. (2010). Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proc. of the 7th Int. Conf. on Language Resources and Evaluation*, pages 2200–2204.

Cambria, E., Speer, R., Havasi, C., and Hussain, A. (2010). Senticnet: A publicly available semantic resource for opinion mining. In *AAAI fall symposium: commonsense knowledge*.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, (20):213–220.

Cruz, G., Huzita, E., and Feltrim, V. (2016). Estimating trust in virtual teams - a framework based on sentiment analysis. In *Proc. of the 18th Int. Conf. on Enterprise Information Systems (ICEIS 2016)*, pages 464–471.

Dodds, P. S. and Danforth, C. M. (2010). Measuring the happiness of large-scale written expression: Songs, blogs, and presidents. *Journal of Happiness Studies*, 11(4):441–456.

Goncalves, P., Benevenuto, F., and Almeida, V. (2013). O que tweets contendo emoticons podem revelar sobre sentimentos coletivos. In *Proc. of the II Brazilian Workshop on Social Network Analysis and Mining (BraSNAM)*, pages 1–12.

Goncalves, P., Dores, W., and Benevenuto, F. (2012). Panas-t: Uma escala psicometrica para analise de sentimentos no twitter. In *Proc. of the I Brazilian Workshop on Social Network Analysis and Mining (BraSNAM)*.

Guzman, E., Azócar, D., and Li, Y. (2014). Sentiment analysis of commit comments in github: An empirical study. In *Proc. of the 11th Working Conf. on Mining Software Repositories*, pages 352–355.

Herbsleb, J. D. and Moitra, D. (2001). Global software development. *IEEE Software*, 18(2):16–20.

Jongeling, R., Datta, S., and Serebrenik, A. (2015). Choosing your weapons: On sentiment analysis tools for software engineering research. In *IEEE Int. Conf. on Software Maintenance and Evolution*, pages 531–535. IEEE.

Liu, B. (2012). *Sentiment Analysis and Opinion Mining*. Morgan & Claypool Publishers.

Mohammad, S. M., Kiritchenko, S., and Zhu, X. (2013). Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets. *Computing Research Repository (CoRR)*, abs/1308.6242.

Murgia, A., Tourani, P., Adams, B., and Ortu, M. (2014). Do developers feel emotions? an exploratory analysis of emotions in software artifacts. In *Proc. of the 11th Working Conf. on Mining Software Repositories*, pages 262–271.

Nielsen, F. Å. (2011). A new anew: Evaluation of a word list for sentiment analysis in microblogs. *Computing Research Repository (CoRR)*, abs/1103.2903.

O'Conchuir, E., Holmstrom, H., Agerfalk, P., and Fitzgerald, B. (2006). Exploring the assumed benefits of global software development. In *Int. Conf. on Global Software Engineering*, pages 159–168.

Park, J., Barash, V., Fink, C., and Cha, M. (2013). Emoticon style: Interpreting differences in emoticons across cultures. In *Proc. of the 7th Int. AAAI Conf. on Weblogs and Social Media*.

Sengupta, B., Chandra, S., and Sinha, V. (2006). A research agenda for distributed software development. In *Int. Conf. on Software Engineering*. ACM.

Sinha, V., Lazar, A., and Sharif, B. (2016). Analyzing developer sentiment in commit logs. In *Proc. of the 13th Int. Conf. on Mining Software Repositories*, pages 520–523.

Thelwall, M. (2013). Heart and soul: Sentiment strength detection in the social web with sentistrength. *Proceedings of the CyberEmotions*, pages 1–14.

Tourani, P., Jiang, Y., and Adams, B. (2014). Monitoring sentiment in open source mailing lists-exploratory study on the apache ecosystem. In *Proc. of the 2014 Conf. of the Center for Advanced Studies on Collaborative Research*, pages 74–95.

Watson, D., Clark, L. A., and Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: the panas scales. *Journal of personality and social psychology*, 54(6):1063.