The Challenges of Sentiment Detection in the Social Programmer Ecosystem

Nicole Novielli, Fabio Calefato, Filippo Lanubile
University of Bari
Dipartimento di Informatica
Bari, Italy
{nicole.novielli, fabio.calefato, filippo.lanubile}@uniba.it

ABSTRACT

A recent research trend has emerged to study the role of affect in in the social programmer ecosystem, by applying sentiment analysis to the content available in sites such as GitHub and Stack Overflow. In this paper, we aim at assessing the suitability of a state-of-the-art sentiment analysis tool, already applied in social computing, for detecting affective expressions in Stack Overflow. We also aim at verifying the construct validity of choosing sentiment polarity and strength as an appropriate way to operationalize affective states in empirical studies on Stack Overflow. Finally, we underline the need to overcome the limitations induced by domain-dependent use of lexicon that may produce unreliable results.

Categories and Subject Descriptors

H.1.2 [**User/Machine Systems**]: Human factors

General Terms

Human Factors.

Keywords

Online Q&A, Technical Forum, Sentiment Analysis, Stack Overflow, Social Programmer, Social Software Engineering

1. INTRODUCTION

Software engineering involves a large amount of social interaction, as programmers often need to cooperate with others, whether directly or indirectly. However, we have become fully aware of the importance of social aspects in software engineering activities only over the last decade. In fact, it was not until the recent diffusion and massive adoption of social media that we could witness the rise of the "social programmer" [41] and the surrounding ecosystem [42].

Social media has deeply influenced the design of software development-oriented tools such as GitHub (i.e., a social coding site) and Stack Overflow (i.e., a community-based question answering site) [43]. Stack Overflow, in particular, is an example of an online community where social programmers do networking by reading and answering others' questions, thus participating in the creation and diffusion of crowdsourced documentation. In our previous work, we argued and proved that among the non-technical factors, which can influence the members of online communities, the emotional style of a technical contribution does affect its probability of success [29], [9]. More specifically, our effort is to understand how expressing affective states in Stack Overflow influences the probability for askers of eliciting an accepted answer and the probability for answerers of having an answer accepted.

Our research follows a recent trend that has emerged to study the role of affect in social computing. For example, Kucuktunc et al. [19] performed a large-scale sentiment analysis study on Yahoo! Answers to assess the impact of the semantic orientation of a post on its perceived quality. Althoff et al. [1] found that expressing gratitude in a question is positively correlated with success of altruistic requests in Reddit.com. Guzman et al. [17] perform sentiment analysis of commit comments in GitHub and demonstrate that a correlation exists between emotions and other factors such as the programming language used in a project, the geographical distribution of the team and the day of the week. Similarly, Guzman and Bruegge [16] used a sentiment analysis tools for detecting the polarity, i.e., the positive or negative semantic orientation of a text, to investigate the role of emotional awareness in software development teams.

What these studies have in common is that they applied sentiment analysis techniques to crowd-generated content relying on polarity as the only dimension to operationalize affect. However, polarity is only one of the possible dimensions of affect, which could be also modeled in terms of its duration, activation, cognitive triggers, and specificity [11]. Still, polarity is the most used dimension because of its ease of measurement and the availability of open source and robust analysis tools. In this paper, we argue that polarity, if employed alone, is insufficient for detecting the sentiments of programmers in a reliable manner. Furthermore, we highlight and discuss the challenges existing when sentiment analysis techniques are employed to assess the affective load of text containing technical lexicon, as typical in the social programmer ecosystem.

The remainder of the paper is structured as follows. In Section 2, we first provide an overview of detecting affective states from text, including a state-of-the-art in the field of sentiment analysis. Then, in Section 3, we perform a qualitative analysis to show the limits of only using polarity to measure the sentiment expressed in questions and answers in Stack Overflow. The findings from our analysis are then discussed in Section 4, where we also outline the future research directions.

SSE'15, September 1, 2015, Bergamo, Italy
© 2015 ACM. 978-1-4503-3818-9/15/09...$15.00
http://dx.doi.org/10.1145/2804381.2804387

2. MINING AFFECTIVE STATES FROM TEXT

2.1 Affect Modeling Theories

Nowadays, affective computing is an established discipline [35] and sensing affective states from text is now regarded as fundamental in several domains, from human-computer interaction [7][21][30] to software engineering [8][26][16].

Affective states vary in their degree of stability and in their ‘object-oriented specificity’ [11], ranging from personality traits to emotions. *Personality traits* are long-standing, organized sets of characteristics of a person that uniquely influences cognitions, motivations and behaviors. *Emotions* are transient and typically complex, episodic, dynamic and structured events, which involve perceptions, thoughts, feelings, bodily changes, and personal dispositions to experience further emotional episodes. Emotions are episodic and dynamic in that, over time, the elements can come and go, depending on all sorts of factors. Other states, such as *interpersonal stance* or *attitudes* are in a middle of this scale: they are initially triggered by individual characteristics like personality, social role, and others but may vary, in valence (i.e. positive vs. negative) and intensity (i.e. low arousal vs. high arousal) by episodes occurring during communication.

Mining affective states from text involves, on one hand, to model them according to bi-dimensional models representing the affect polarity or valence and its level of activation or intensity; on the other hand, some studies explicitly deal with discrete emotion labeling of text, by looking for linguistic cues of specific affective states. Psychologists worked at decoding emotions for decades, by focusing on two main questions: (i) how can emotions be classified? (ii) What is their functioning?, i.e., How are they triggered? How do they affect behavior? Which is the role played by cognition? Two points of view prevailed: the first one assumes that a limited set of basic emotions exists, while the second one consider emotions as a continuous function of one or more dimensions. It is the case of the ‘circumplex model’ of affect [40], which models emotions along a bi-dimensional representation schema, including valence (pleasant vs. unpleasant) and arousal (activation vs. deactivation) of emotions. Conversely, theories following the discrete trend agree on the idea that a limited set of basic emotions exists, although consensus about the nature and the number of these basic emotions has not been reached. Ekman defines a basic emotion as having specific feelings, universal signals and corresponding physiological changes [12]; Lazarus describes nine negative (Anger, Anxiety, Guilt, Shame, Sadness, Envy, Jealousy, Disgust) and six positive (Joy, Pride, Love, Relief, Hope, Compassion) emotions, with their appraisal patterns: positive emotions are triggered if the situation is congruent with one of the individual’s goals; otherwise, negative emotions are triggered [20]; Plutchik defines discrete emotions as corresponding to specific adaptive processes: reproduction, safety, etc. [36]. The Plutchik’s model identifies eight primary emotions (i.e., anger, fear, sadness, disgust, surprise, anticipation, trust, and joy), graphically depicted as a wheel (see Figure 1). Opposite affective states, with respect to both valence and intensity, are placed opposite of each other using complementary colors. Strength of emotions increases towards the center of the model, with low activation affective states lying next to the circumference.

During the last decade, in computational linguistics several approaches have been investigated for mining affect from text. With respect to the specific goals addressed and to the adopted theory, researchers model and detect affect at a different granularity. Several studies refer to discrete emotion categorization. For example, Liu et al. [23] propose a method based on large-scale real-world knowledge about the way people usually make appraisals of everyday situations. The approach exploits generic knowledge basis of commonsense to identify the six Ekman’s basic emotional states (happy, sad, angry, fearful, disgusted, and surprised) through the text analysis. Neviarouskaya et al. [27] use a rule-based approach that includes consideration of the deep syntactic structures for emotion computation in the text. Experiments are performed on different corpora to identify nine emotional labels (plus the neutral one). Compared with other state-of-the-art techniques, the method shows promising results in fine-grained emotion recognition. Other studies, rather focuses on the valence (i.e. the positive or negative orientation) of affective states conveyed in natural language interactions. It is the case of Litman et al. [22], who defined an annotation scheme to label emotions in tutoring dialogs along a linear scale (negative, neutral, positive) in order to detect students’ boredom and frustration. Analogously, Batliner et al. [7], define a method for automatically detecting emotionally critical phases in dialogues with customers of an automatic call-center, in order to enhance customers’ satisfaction by adapting the interaction accordingly. Le Tallec et al. [21] studied how to classify emotions in speech by considering the language of hospitalized children interacting with companion robots. Linguistic clues are considered, achieving good results in detecting emotional valence of utterances. Novielli et al. [30], exploit linguistic cues to detect cold vs. warm social attitude of users toward an Embodied Conversational Agents in the domain of simulation of persuasion dialogues.

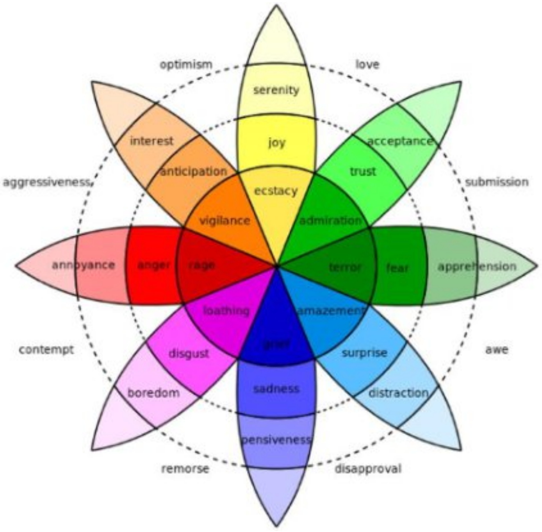

Figure 1. The Plutchik Wheel [36]

2.2 State of the Art on Sentiment Analysis

As far as affect polarity is concerned, sentiment analysis provides researchers with a suite of methods and linguistic resources, which can be exploited for recognizing the semantic orientation of texts. Sentiment analysis is the study of the subjectivity and polarity of a text [33]. More recently, researcher in this field started to deal also with sentiment strength detection, obtaining promising results [47]. Sentiment analysis techniques have been successfully applied to the problem of detecting the valence of affective states conveyed by a text [16], for modeling socioeconomic phenomena [32], and to automatically analyze text for opinion mining purposes [18][33].

2.2.1 Approaches

Traditional approaches to sentiment analysis treat the subjectivity and polarity detection as text classification problems and exploit machine learning algorithms for extracting features to train supervised classifiers on human-annotated corpora. The features employed are typically words (i.e., tokens, stems, lemmata) and part-of-speech tags, also combined in n-grams, that is sequences of *n* contiguous words. Such approaches mainly rely on state-of-the-art machine learning algorithms, such as Support Vector Machines [48], and might also be improved by performing intelligent feature selection [38]. With the worldwide diffusion of social media, a huge amount of textual data has been made available, thus attracting the interest of researchers in this domain [39]. Sentiment analysis on such informal texts poses new challenges due to the presence of slang, misspelled words, hashtags, and links, thus inducing researchers to define novel approaches that include consideration of micro-blogging features [6][25].

However, supervised approaches present the main drawback of being highly domain-dependent. This means that classification models are very likely to perform poorly outside the domain they were trained on [15]. In fact, when training classification models, it is very likely to include consideration of terms that associate with sentiment because of the domain. It is the case of political debates, where names of countries afflicted by wars might be associated to negative sentiments; analogous problems might be observed for the technology domain, where killer features usually referred in positive customers' review usually become obsolete in relatively short periods of time. Such terms are usually referred by researchers as *indirect affective words* to distinguish them from *direct* ones [45]. Indeed, according to the emotion theory defined by Clore et al. [10], it is possible to distinguish between words that directly refer to emotional states (e.g., 'fear', 'joy', 'cheerful', 'sad') and those having only an indirect reference to an emotional state, depending on the context (e.g., the words which indicates emotional causes such as 'killer' or 'monster' or emotional responses to an event such as 'cry' or 'laugh').

To overcome these limitations, lexical approaches are adopted, which exploit the prior sentiment polarity of words in a text, based on lexical resources, i.e. large lexicons of words annotated with their prior polarity (i.e. positive or negative semantic orientation of the word). The overall sentiment of a text is computed based on the prior polarity of the words composing it [28][46] as well as on their contextual polarity (i.e. the polarity conveyed by a word with respect to its context of usage) [4][49]. In fact, words with negative prior polarity can be used to express positive sentiment (as in '*I feel so sorry for you*' where a positive attitude towards the interlocutor is expressed) or even in a neutral context. Furthermore, the effect of contextual valence shifter [37], such as intensifiers or negation, needs to be taken into account since they might intensify, mitigate or even invert the polarity of the word they are associated with. Therefore, lexicon-based approaches are usually integrated with other knowledge such as semantic rules [27][46] or features specific of the communication medium, as for example emoticon lists [47].

2.2.2 Linguistic Resources

In this section, we provide an overview on the state-of-the-art linguistic resources for sentiment analysis and affect mining from text.

Sentiment lexicons are basically organized as lists of words with scores indicating their prior polarity and, in some cases, also the sentiment strength. These lexicons can be differentiated based on how they represent the information about prior polarity of words.

The Bing Liu Lexicon of Opinion Words [18] is a manually created lexicon of about 6800 words, built upon e-commerce customer reviews. The annotation is binary and simply states if a word expresses positive or negative sentiment.

Similarly, the MPQA Lexicon [49] provides a broader list of positive and negative terms (about 8k), also including information about the strength of the sentiment conveyed (i.e. a word can be categorized as either 'weakly' or 'strongly' subjective.The part of speech is also reported, since different prior polarity annotations can be assigned to the same word, based on the role it plays in the discourse. The MPQA Lexicon is part of the OpinionFinder[1], a system for automatic identification of subjective sentences in documents.

The NRC Emotion Lexicon [24] contains about 14K entries consisting of words for which annotation is provided based on both polarity and a set of discrete emotion labels (i.e. anger, fear, sadness, disgust, surprise, anticipation, trust, and joy). The lexicon has been created starting from an annotation of word-sense pairs. Each word-sense pair is annotated by at least three annotators and the final word-level lexicon was created by considering the union of the emotion labels provided for all the senses of a word.

Similarly, the NRC Hashtag Sentiment Lexicon and the Sentiment140 Lexicon provide lists of words with their sentiment association score, calculated as pointwise mutual information with respect to collections of positive and negative tweets [25]. Unlike the previous ones, these two lexica have been created exploiting a completely automatic training procedure based on a corpus of English tweets annotated using positive and negative hashtags and emoticons as 'noisy' labels.

Another widely used lexicon designed on purpose for serving sentiment analysis tasks is SentiWordNet 3.0 [13]. SentiWordNet extends Word-Net [14] by associating positive, negative and objective scores to each synset (i.e., set of synonyms), where the three scores sum up to 1. A word can receive multiple polarity scores if it occurs in more than one synset (see example scores for the word 'good' reported in Table 1). Thanks to the availability of explicit objective scores, additional features can be computed to model the presence of neutral terms, as reported in [6].

WordNet Affect [44] also extends the WordNet database with affective labels (a-labels) for annotation of synsets. One or more a-labels may be assigned to a synset. The resource also includes a-labels representing moods, situations eliciting emotions, or emotional responses (see examples in Table 2).

In this review, we include the Linguistic Inquiry and Word Count (LIWC) taxonomy, even if it has been developed in the scope of broader psycholinguistic research [34] without being explicitly designed for sentiment analysis. LIWC organizes words into psychologically meaningful categories based on the assumption that words and language reflect most part of cognitive and emotional phenomena involved in communication. Previous research has shown how the language use varies with respect to the communicative intention, thus making possible to distinguish between objective and subjective statements as well as between agreement and disagreement expressions [31]. In fact, among the

[1] http://mpqa.cs.pitt.edu/opinionfinder

word classes included in the taxonomy, LIWC provides linguistic categories that draw distinctions between negative and positive emotion lexicon, which research could use to derive word-count metrics to assess the affective load of a text.

Table 1. SentiWordNet scores for the word 'good'

Scores	Sense ID and gloss
as an adjective	

Positive = 0.75
Negative = 0
Objective = 0.25
	good#1
	Having desirable or positive qualities especially those suitable for a thing specified (as in 'a good joke')

Positive = 0
Negative = 0
Objective = 1
	good#2
	having the normally expected amount (as in 'gives good measure')

Positive = 1
Negative = 0
Objective = 0
	good#6
	agreeable or pleasing
	(as in 'we all had a good time')

as a noun	

Positive = 0.5
Negative = 0
Objective = 0.5
	good#1
	benefit (as in 'for your own good')

Positive = 0.875
Negative = 0
Objective = 0.125
	good#2
	moral excellence or admirableness (as in 'there is good in people')

Positive = 0
Negative = 0
Objective = 1
	good#4
	commodity, article of commerce

Table 2. A-Labels in WordNet Affect with Examples

A-label	Example of Synsets
EMOTION	noun 'anger', verb 'fear'
MOOD	noun 'animosity', adjective 'fear'
TRAIT	noun 'aggressiveness', adj. 'competitive'
COGNITIVE State	noun 'confusion', adj. 'dazed'
PHYSICAL State	noun 'illness'
HEDONIC Signal	noun 'hurt', noun 'suffering'
Emotion-eliciting SITUATION	noun 'awkwardness'
Emotional RESPONSE	noun 'cold sweat', verb 'tremble'
BEHAVIOR	noun 'offense', adj. 'inhibited'
ATTITUDE	noun 'intolerance', noun 'defensive'
SENSATION	noun 'coldness', noun 'feel'

3. ANALYSIS OF AFFECTIVE STATES IN STACK OVERFLOW

3.1 Dataset and Instrumentation

We built the dataset for our analysis starting from the official Stack Overflow dumps.[2] As for the questions and askers' comments, they were extracted from the Stack Overflow dump, updated on May 2014. Instead, the answers and answerers' comments were extracted from the official Stack Overflow dump of September 2014, as described in [9]. For each question, answer and comment in our dataset, we annotated the sentiment score, measured in terms of text polarity and its strength. Then, we built our final dataset for the qualitative analysis by opportunistically choosing the cases with the highest sentiment score.

In order to measure the sentiment load of a contribution, we look for affective lexicon in the body of questions, answers and comments. Specifically, we measure the overall positive/negative polarity of a text as well as the sentiment strength. We deliberately choose to capture the *sentiment* of text, that is, its positive/negative semantic orientation of the text. Sentiment is calculated for each post in our dataset using SentiStrength[3], a state of the art tool already employed in previous research on sentiment analysis in social computing [1][17][19], which has been designed to overcome the limitations due to domain-dependence of supervised approaches. SentiStrength is a lexicon-based classifier that exploits rules and a lexicon built by combining entries from different linguistic resources. Furthermore, it has been validated for the social web and therefore it is robust for the analysis of informal text including web jargon (such as emoticons or abbreviations) [47].

Based on the assumption that a sentence can convey mixed sentiment, SentiStrength outputs both positive and negative sentiment scores for any input text written in English. It assigns the overall positive and negative scores to a text by considering the maximum among all the sentence scores**.** In SentiStrength, positive sentiment scores range from ±1 (neutral) to +5 (extremely positive/negative). In our analysis, we adjust the sentiment score and map them into the ±[0,4] interval, with zero indicating the absence of positive or negative sentiment ('neutral' texts). These metrics represent the overall positive/negative polarity and strength of the sentiment conveyed by a post, whether a question, an answer or a comment. For each post, we issue both its positive and negative score. When both are null, the post is considered as *neutral* (no sentiment is expressed through the lexicon employed). A few examples are provided in Table 3.

We analyzed the top 100 questions and follow-up askers' comments with the highest positive and negative scores (400 cases overall). Analogously, we analyzed the 100 top answers and follow-up answerers' comments with the highest positive and negatives score (400 additional cases).

[2] https://archive.org/details/stackexchange

[3] http://sentistrength.wlv.ac.uk

Table 3. Examples of sentiment expression in questions in Stack Overflow with associated sentiment scores

"*I have very simple and* ***stupid trouble*** *[...] I'm* ***pretty confused****, explain please, what is wrong*?" Positive Score: 0; Negative Score: 1
*"****Thank you****, that was* ***really helpful****"* Positive score: 3; Negative Score: 0

3.2 Affect in Questions and Askers' Comments

As for questions with an overall negative polarity, the analysis reveals that the askers generally express a negative emotion associated with their technical issue or report a negative opinion. In the first case, information seekers mainly express their frustration for not being able to solve a problem as in the following examples:

- *"I am not sure what I did in a previous life to warrant this, it must have been bad! I am getting buried in a world of xml [...]"*
- *"I recently got stuck on an odd problem"*.

Conversely, negative opinions mainly refer to a preference or evaluation about a technical issue, as in these examples.

- *"They use it to clean up connections, which is really scary"*
- *"It is very painful to add multiple tickets to Trac",*
- *"I find this incredibly annoying with Dreamweaver"*.

As for negative comments added by the asker, opinions about tools and resources are the main reason for use of negative affective lexicons as in

- *"I really hate those properties panels that don't look the same whether they are VB/C# winform/web. This sucks!"*

Again, we notice information seekers expressing their distress:

- *"This is driving me nutz :-("*

frustration:

- *"Unfortunately Xcode doesn't use CodeSense for editing files outside a project; which is incredibly frustrating."*
- *"I still get 400 bad request page! :(("*

and fear:

- *"there's no way to do this I'm afraid :("*

Cases of humor are also detected:

- *"Haha that comment made me laugh my heart out. Makes me kind of proud of how horrible my code can be!"*
- *"I would also like to add that I do find humor in these sorts of problems. This is because the only other option is to become horribly horribly angry. And who wants that?!"*

as well as a few instances of offending sarcastic comments:

- *"Do u know [the] answer ? If not .... then u might know how hard is homework. Ignorance is bliss"*.

As for questions with positive sentiment score, we do not observe actual emotion reporting. On the contrary, we rather find opinions, as in the following examples:

- *"Given past frustrations with Win32 inextensibility, this seems like a good thing"*
- *"[...] I'm checkin out what WCF has to offer. It seems very flexible and a great next step up from"*.
- *"I'm loving the built-in Clang!! Kudos to Xcode team!")*.

Furthermore, we observe a considerable use of affective lexicon for politeness expressions. This is typical of so-called '*behabitives*' [5], speech acts in which no real feelings are expressed but still emotional words are employed to convey other communicative intentions. A typical use found in our Stack Overflow dataset is in expressing gratitude in advance towards potential helpers, as in the following examples:

- *"Any help is hugely appreciated!"*
- *"Any example would be super magnificent!"*

Evidence of analogous use of positive lexicon is found in comments. We found cases where no true feelings experienced by the asker are reported. Rather, gratitude is expressed (e.g., *"Wow, great! Thanks"*) as well as appreciation for the solution provided in the previous comments by other Stack Overflow users, e.g.:

- *"Excellent link! You should have posted it as an answer :P"*
- *"Excellent! Thanks for the info."*

Technical opinions are also reported in a few cases:

- *"As modeler we use Activity Designer. Excellent tool!"*

3.3 Affect in Answers and Answerers' Comments

As for answers, we observe that a negative lexicon is used for expressing actual emotions in the great majority of cases. However, the negative polarity of these answers does not involve a negative judgment of the asker but it rather results in either an attempt of showing empathy towards the asker as in the following examples:

- *"If you are really worried about storage usage [...]"*
- *"This could be very annoying but it is simply solved"*
- *"[...] This will make your experience a lot less frustrating"*

or in a criticism towards a technological issue, as in the following cases:

- *"This could work, but feels really awful"*
- *"This is extremely ugly for loop construction."*

As for comments to answers, we found they are very rich in affective lexicon too. We observe that Stack Overflow users express a wide variety of affective states in comments. We speculate that this might occur because comments are seen as a 'free zone' since reputation mechanisms do not apply to comments. More in detail, as for positive comments we found that the main affective states are gratitude, as in

- *"Thanks for the feedback, it was a pleasure!"*

wishes, as in

- *"Happy coding!"*

and positive feelings linked to satisfaction and happiness for the help provided, as in the following example

- Asker*: "Thanks, that helped. Case closed!"*- Answerer*: "Thanks for the feedback! It was a pleasure!"*

As for negative comments, we observe a wider variety of emotions. In some cases the negative polarity of lexicon actually conveys a negative attitude towards the interlocutor (in this case, the author of the question)

- *“Didn't notice the horrid inline jQuery”*
- *“Added some instructions for the really hopeless cases”*
- *“Arrrghhh, how I hate those people who downvote answers without leaving a comment as for why the downvote…”*

Conversely, negative lexicon may also be used to convey a positive attitude towards the reader. In such cases, the writer attempts to convey signal of empathy towards the interlocutor, as in the following examples, where people apologize for not being able to provide further help:

- *“To explain my regrettably unfriendly comment (sorry about that).”*
- *“You could try this one (not optimal, I am afraid)”*
- *“I’m afraid I can’t help you any further with this issue!”*

Furthermore, users employ a negative lexicon also for expressing opinions on controversial technical issues, as in:

- Asker*: “But what if you do have to worry about spaces in your filenames?”* - Answerer*: “Then you've got major problems! Let me meditate on it; it is extremely unpleasant, whatever.”*
- *“Sorry for all the editing but this is a ridiculously complex issue”*

3.4 Domain-dependence of Sentiment Analysis

Due to the presence of domain-specific lexicon, examples of false positives in negative sentiment detection emerge from the analysis of both comments and answers, such as in

- *“You are vulnerable to this bug”*
- *“What is the best way to kill a critical process”*
- *“I am missing a parenthesis. But where?”*

As already highlighted in Section 2, the domain-dependency of sentiment analysis tools is a known problem [15] meaning that applying a tool outside the domain in which it was tested, may produce unreliable results.

In the case of Stack Overflow, this bias is emphasized by the fact that a social Q&A site is explicitly designed for people looking for help because in trouble with a domain-specific issue. Therefore, discussion on Q&A sites are intrinsically skewed towards negative polarity because they are naturally rich in ‘problem’ lexicon, which does not necessarily indicate the intention to show any affective state, as in the following example, for which a negative score is inaccurately issued by SentiStrength:

- *“I have a trouble […]. Please, explain what is wrong”*

4. DISCUSSION AND FUTURE WORK

The findings of our qualitative investigation highlight open challenges for sentiment analysis in social software engineering and inspire directions for future work. First, in contrast with the current trend of using sentiment to operationalize the affect dimension in empirical studies, our analysis advocates in favor of the adoption of more appropriate affective state models. The wide variety of emotions, attitudes and opinions retrieved in our dataset confirm that affect is a quite complex phenomenon whose polarity is only one dimension of analysis.

Even when performed with state-of-the-art tools, measuring the polarity of a text is not sufficient to capture the attitude of the sender towards the recipient of a text, despite the use of affective loaded lexicon which may refer to technical issue rather than being addressed to the interlocutor.

Furthermore, the wide variety of affective states expressed in Stack Overflow posts recommends a more fine-grained investigation of the role of emotions. Depending on the specific goals addressed, researchers could be interested in issuing a discrete label describing the affective state expressed (e.g., frustration, anger, sadness, joy, satisfaction) as different affective states may be relevant to different context of interaction and tasks. For example, being able to identify harsh comments towards technical matters could be useful in detecting particularly challenging questions that have not been exhaustively answered, which is a goal addressed by current research on effective knowledge-sharing in Stack Overflow [2]. Conversely, detecting attitude towards the interlocutor could be of help for the community moderators that could intervene in order to guide the users’ behavior towards a more constructive pattern of interaction towards cooperative problem solving. Indeed, previous research on success of questions has shown how strong negative emotions in follow-up discussions discourage participation [3]. In fact, even if Stack Overflow guidelines include a ‘Be nice’ section, in which users are invited to be patient and avoid offensive behavior, people might not be prepared for effectively dealing with the barriers of social media to non-verbal communication. This clearly emerges as an open problem in the Stack Exchange community as discussed by users, which complain about harsh comments mainly coming from expert contributors.

Finally, we underline the need for tuning state-of-the-art resources for sentiment detection. Indeed, an open challenge for sentiment analysis is to overcome the limitations induced by domain-dependent use of lexicon. This is consistent with Wittgenstein’s *meaning-is-use* assumption [50], claiming that the meaning of an expression is fully determined by its use.

In our future work, we will improve lexicon-based sentiment analysis approaches and fine-tune state-of-the-art resources by exploiting semantic features [6] for appropriately dealing with domain-dependent use of lexicon, in order to distinguish accurately neutral sentences from emotionally loaded ones in Stack Overflow discussions.

5. ACKNOWLEDGMENTS

This work is partially funded by the ATS Romantic under the Apulian ICT Living Labs program and by the project ‘Investigating the Role of Emotions in Online Question & Answer Sites’, funded by MIUR (Ministero dell’Università e della Ricerca) under the program "Scientific Independence of young Researchers" (SIR).

6. REFERENCES

[1] Althoff, T., Danescu-Niculescu-Mizil, C., and Jurafsky, D. 2014. How to Ask for a Favor: A Case Study on the Success of Altruistic Requests. *In Proceedings of the 8th International AAAI Conference on Weblogs and Social Media (ICWSM 2014)*.

[2] Anderson, A., Huttenlocher, D., Kleinberg, J. and Leskovec. J. 2012. Discovering value from community activity on focused question answering sites: a case study of stack overflow. *Proc. of KDD '12*. ACM, 850-858.

[3] Asaduzzaman, M., Mashiyat, A.S., Roy, C.K., Schneider, K.A. 2013. Answering questions about unanswered questions of Stack Overflow, *Proc. of MSR 2013*, 97-100.B.

[4] Akkaya, C., Wiebe, J., Conrad, A., and Mihalcea, R. 2011. Improving the Impact of Subjectivity Word Sense Disambiguation on Contextual Opinion Analysis. *Proc. Of CoNNL '11*, 87-96.

[5] Austin, J 1962. *How to do Things with Words*. Oxford University Press, New York.

[6] Basile, P. and Novielli, N. 2015. UNIBA: Sentiment Analysis of English Tweets Combining Micro-blogging, Lexicon and Semantic Features. *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, ACL, 595—600.

[7] Batliner, A., Fisher, K., Huber, R., Spilker, J., No¨th, E. 2003. How to find trouble in communication. Speech Communication 40, 117–143.

[8] Brooks, M., Kuksenok, K., Torkildson, M.K., Perry, D., Robinson, J.J., Scott, T.J., Anicello, O., Zukowski, A., Harris, P., and Aragon, C.R. 2013. Statistical affect detection in collaborative chat. In *Proceedings of the 2013 conference on Computer supported cooperative work* (CSCW '13). ACM, New York, NY, USA, 317-328.

[9] Calefato, F., Lanubile, F., Marasciulo, M.C., Novielli, N. 2015. MSR Challenge: "Mining Successful Answers in Stack Overflow." In *Proc. 12th IEEE Working Conf. on Mining Software Repositories* (MSR '15), Florence, Italy, May 16-17, 2015.

[10] Clore, Gerald L.; Ortony, Andrew; Foss, Mark A. 1987. The psychological foundations of the affective lexicon. *Journal of Personality and Social Psychology*, Vol 53(4), Oct 1987, 751-766.

[11] Cowie, R. 2006. *Emotional life, terminological and conceptual clarifications*. Available at the HUMAINE Portal: http://emotion-research.net/deliverables/D3i final.pdf

[12] Ekman, P. 1999. *Basic Emotions. Handbook of Cognition and Emotion*, John Wiley & Sons Ltd.

[13] Esuli, A. and Sebastiani, F. 2006. Sentiwordnet: A publicly available lexical resource for opinion mining. *Proceedings of LREC*, volume 6, 417–422.

[14] Fellbaum C. WordNet: An Electronic Lexical Database. *Language, Speech, and Communication*. MIT Press, 1998.

[15] Gamon, M., Aue, A., Corston-Oliver, S. and Ringger, E. 2005. Pulse: Mining customer opinions from free text. *LNCS 3646*, 121-132.

[16] Guzman, E. and Bruegge, B. 2013. Towards emotional awareness in software development teams. *In Proceedings of the 2013 9th Joint Meeting on Foundations of Software Engineering (ESEC/FSE 2013)*. ACM, New York, NY, USA, 671-674.

[17] 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 (MSR 2014)*. ACM, New York, NY, USA, 352-355.

[18] Hu, M. and Liu, B. 2004. Mining and Summarizing Customer Reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 168–177.

[19] Kucuktunc, O., Cambazoglu, B.B., Weber, I., and Ferhatosmanoglu, H. 2012. A large-scale sentiment analysis for Yahoo! answers. *In Proceedings of the fifth ACM international conference on Web search and data mining (WSDM '12)*. ACM, New York, NY, USA, 633-642.

[20] Lazarus R S (1991) *Emotion and adaptation*. New York: Oxford University Press.

[21] Le Tallec, M., Antoine, J.-Y., Villaneau, J. and Duhaut, D. 2011. Affective interaction with a companion robot for hospitalized children: a linguistically based model for emotion detection. *Proc. of the 5th Language and Technology Conference (LTC2011)*.

[22] Litman, D. Forbes-Riley, K., Silliman, S. 2003. Towards emotion prediction in spoken tutoring dialogues. In: *Companion Proceedings of the Human Language Technology Conference: 3rd Meeting of the North American Chapter of the Association for Computational Linguistics (HLT/NAACL)*, Kluwer, Amsterdam, pp. 52–54.

[23] Liu, H.,Lieberman, H. and Selker, T. 2003. A model of textual affect sensing using real-world knowledge. *Proceedings of the 8th international conference on Intelligent user interfaces*, ser. IUI

[24] Saif M. Mohammad and Peter D. Turney. 2010. Emotions Evoked by Common Words and Phrases: Using Mechanical Turk to Create an Emotion Lexicon. In *Proc. of the NAACL HLT 2010 Workshop CAAGET '10*, pages 26–34.

[25] Saif Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. 2013. Nrc-canada: Building the state-of-the art in sentiment analysis of tweets. *Proc. SemEval 2013*, 321–327.

[26] 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 Working Conference on Mining Software Repositories (MSR 2014)*. ACM, New York, NY, USA, 262-271.

[27] Neviarouskaya, A., Prendinger, H. and Ishizuka, M. 2011. Affect analysis model: Novel rule-based approach to affect sensing from text. *Nat. Lang. Eng.*, vol. 17(1), 95–135.

[28] Nielsen, F.A. 2011. A New ANEW: Evaluation of a Word List for Sentiment Analysis in Microblogs. *MSM 2011*: 93-98.

[29] Novielli, N., Calefato, F., Lanubile, F. 2014. Towards discovering the role of emotions in stack overflow. In *Proc. of 6th Int'l Workshop on Social Software Engineering* (SSE '14). ACM, New York, NY, USA, 33-36

[30] Novielli, N., de Rosis, F. and Mazzotta, I. 2010. User Attitude Towards an Embodied Conversational Agent: Effects of the Interaction Mode. *Journal of Pragmatics*, 42(9), Elsevier Science, 2385-2397.

[31] Novielli N. and Strapparava, C. 2013. The Role of Affect Analysis in Dialogue Act Identification. *IEEE Transactions on Affective Computing*, 4:439–451.

[32] O'Connor, B, Balasubramanyan, R., Routledge, B., and Smith, N. 2010. From tweets to polls: Linking text sentiment to public opinion time series. In *Intl AAAI Conf. on Weblogs and Social Media (ICWSM)*, volume 11, pages 122–129.

[33] Pang, B. and Lee, L. 2008. Opinion Mining and Sentiment Analysis. *Foundations and trends in information retrieval*, 2(1-2):1–135.

[34] Pennebaker, J. and Francis, M. 2001. *Linguistic Inquiry and Word Count: LIWC*. Erlbaum Publishers.

[35] Picard, R. W. 2000. *Affective Computing*. MIT Press.

[36] Plutchick, R. 1984. Emotions: A general psychoevolutionary theory. In K.R. Scherer & P. Ekman (Eds) *Approaches to emotion. Hillsdale*, NJ; Lawrence Ealrbaum Associates.

[37] Polanyi, L. and Zaenen, A. 2006. Contextual Valence Shifters. In *Computing Attitude and Affect in Text: Theory and Applications*, The Information Retrieval Series, 1-10

[38] Riloff, E., Patwardhan, S., & Wiebe, J. 2006. Feature subsumption for opinion analysis. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 440-448.

[39] Rosenthal, R., Nakov, P., Kiritchenko, S., Mohammad, S.M., Ritter, A., and Stoyanov, V. 2015. SemEval-2015 Task 10: Sentiment Analysis in Twitter. *In Proceedings of SemEval 2015*.

[40] Russell, J A (2003) *Core affect and the psychological construction of emotion*.

[41] Storey, M-A. 2012. The evolution of the social programmer. In *Proc of the 9th IEEE Working Conf on Mining Software Repositories* (MSR '12). IEEE Press, Piscataway, NJ, USA, 140-140.

[42] Leif Singer, Fernando Figueira Filho, Brendan Cleary, Christoph Treude, Margaret-Anne Storey, and Kurt Schneider. 2013. Mutual assessment in the social programmer ecosystem: an empirical investigation of developer profile aggregators. In *Proceedings of the 2013 conference on Computer supported cooperative work* (CSCW '13). ACM, New York, NY, USA, 103-116.

[43] Storey, M-A., Singer, L., Cleary, B., Figueira Filho, F. and Zagalsky, A. 2014. The (R) Evolution of social media in software engineering. In *Proceedings of the on Future of Software Engineering* (FOSE '14). ACM, New York, NY, USA, 100-116. DOI=10.1145/2593882.2593887

[44] Strapparava, C., and Valitutti, A. 2004. WordNet-affect: an affective extension of WordNet. *In Proceedings of the 4th International Conference on Language Resources and Evaluation*, 1083-1086.

[45] Strapparava, C., Valitutti, A., & Stock, O. 2006. The affective weight of lexicon. *Proceedings of the Fifth International Conference on Language Resources and Evaluation, LREC 2006.*

[46] Taboada, M., Brooke, J., Tofiloski, M., Voll, K.and Stede, M. 2011. Lexicon-based methods for sentiment analysis. *Comput. Linguist.* 37, 2, 267-307.

[47] Thelwall, M. Buckley, K. and Paltoglou, G. 2012. Sentiment Strength Detection for the Social Web. *Journal of the American Society for Information Science and Technology*, 63(1):163-173.

[48] Vapnik, V. 1995. *The Nature of Statistical Learning Theory*. Springer-Verlag.

[49] Wilson, T., Wiebe, J. and Hoffman, P. 2005. Recognizing Contextual Polarity in Phrase-level Sentiment Analysis. *Proc. of HLT '05*, 347–354.

[50] Wittgenstein, L. 1965. *Philosophical Investigations*. The Macmillan Company, New York, NY, USA.