Sentiment Analysis on Stack Overflow Programming Questions

SUMIT SINGH -11713099

Lovely Professional university, Punjab

Sumitpaliya9087@gmail.com

**Abstract**

The sentiment expressed in software engineering (SE) texts has been shown to affect both the productivity and the quality of collaborative work. This is one reason for why sentiment analysis on SE texts has gained attention in research in recent yerars. A large and open resource of SE texts is Stack Overflow (SO). SO is the largest question and answer (Q&A) web site in the Stack Exchange network, and has been the subject for several sentiment analysis studies. It has lately been established that sentiment analyzers trained on social media perform poorly on SE texts, which could challenge the credibility of some of these studies. The Senti4SD sentiment polarity classifier was developed and trained on SO documents to address some of these issues. In this study, random samples of SO documents are drawn and then classified with Senti4SD. The classification into positive, negative and neutral sentiment is used to model the sentiment probability distributions of different document types on SO as a whole, as well as for the eight most popular programming languages. The results indicate that the sentiment of a document is correlated to both the document type and the associated programming language. Among the three sentiment classes, neutral sentiment dominates throughout all SO documents. However, the reliability of the results are reduced by concerns regarding the accuracy of Senti4SD, vaguely specified pre-processing steps and possibly varying classifier bias in different subdomains. In conclusion, further research on sentiment classifiers for SE is needed before any detailed comparative studies of this kind can yield reliable results.

**Introduction**

Stack Overflow (SO) is the largest question and answer (Q&A) website in the Stack Exchange network, and deals solely with programming questions. A programmer who is unable to solve a problem can post a question, which can then be answered by other programmers. Questions and answers (collectively referred to as posts) can be commented, commonly to ask for clarification of a question or to point out errors in an answer [1], [2]. Answering a question is a collaborative effort of the community, as answers are posted, commented and subsequently edited to pin down the solution(s). It is common for authors of questions and answers to edit their posts after receiving feedback in comments, but expert users can circumvent this collaborative aspect and directly edit other users’ posts [3]. As of March 25th 2018, SO has more than 8.6 million users, 16 million asked questions and 24 million answers [4]. The large amount of information contained in the posts on SO makes it a valuable resource for beginning and professional programmers alike. Questions are marked with 1-5 tags specifying which fields they relate to [5]. For example, a question asking about linked lists in Python may be tagged with python and linked-list, making it easy to find questions on a per-field basis.

**Sentiment on SO**

Sentiment in the workplace has been shown to be an important factor for both productivity and quality in collaborative work [6], [7], [8]. Considering SO as a collaborative place of work, the sentiment expressed in posts and comments could affect the efficiency of the site. Previous research has shown that user interaction on SO is not conducted in an entirely neutral fashion, but rather contains notable amounts of both positive and negative sentiment [9], [10], [11]. There is also evidence to suggest that SO is sometimes perceived as a hostile environment, as indicated by several notable Stack Exchange posts [12], [13], [14], the 2014-15 overhaul of the Stack Exchange Be Nice policy [15] and a recent post on the official Stack Overflow blog [16].

**RQ1: Does sentiment expressed in SO documents depend on document type?**

Because of SO’s size, there are many sub-communities within the site, and sentiment distribution could differ among these. A natural categorization into sub-communities is by programming language. A study on GitHub commit messages found that sentiment differed among programming languages [17]. Another study conducted on SO posts showed that politeness (which is related to sentiment) was also dependent on programming language [18]. This leads to the second and final research question of this study:

**RQ2: Does sentiment expressed in SO documents depend on the associated programming language?**

In this study, we use the Senti4SD sentiment polarity classifier1 to classify a total of almost 800,000 randomly sampled questions, answers and comments from 2017, gathered from the Stack Exchange Data Dump [19]. The results are then used to produce estimations of the sentiment probability distributions of different document types for SO as a whole, as well as for the 8 most popular programming languages.

**Background**

**Natural Language Processing**

Natural Language Processing (NLP) is the extraction of meaning from human language by way of computation [20]. In the domain of spoken language, NLP comes in the form of automatic speech recognition (ASR) [21]. Commercial applications have already come a long way in this field [22]. However, it is still an active field of research, especially in terms of developing neural network-based solutions [23, pp. 1–9]. NLP as been a field of research for written language since the 1950’s [22]. Possible applications are many, including text-to-speech, automatic translation and information retrieval. NLP has historically been approached in one of two ways: with our without statistics (including machine learning) [20]. Because of hardware limitations, early approaches primarily consisted of hand-crafted, rule-based solutions. However, the complexity of natural languages heavily challenged pure rule-based systems [22]. Statistical methods were later adopted to improve the performance of NLP for written language, and in this, the ASR community was the driving force [24]. One interesting subfield of NLP for written language, which is also the main focus of this report, is sentiment analysis of written text.

**Sentiment Analysis**

The increased popularity of machine learning methods and the large datasets provided by the World Wide Web in the 21st century have led to a boost in research related to sentiment analysis. Sentiment analysis, or opinion mining, is the study of opinion using computational methods [25]. More formally, it can be defined as the discovery of sentiment expressed by an opinion holder towards some aspect of an entity in an opinionated document [26], [27]. Document-level sentiment classification treats a multisentence document as a single entity, while sentence-level sentiment classification classifies sentiment on a per-sentence basis [25]. Sentiment can be classified based on its polarity (positive, negative or neutral), quantized according to some given scale or categorized into different emotions such as anger and joy [26], [28]. There are many possible application areas for sentiment analysis, including product review understanding, recommendation systems and question answering services [25]. Sentiment analysis methods can be broadly divided into two categories: the machine learning approach and the lexical-based approach [28]. The machine learning approach can be further divided into supervised and unsupervised methods, with the former requiring labeled training data guiding the classification, and the latter discovering groups of similar data without the provision of ground-truth [29]. The challenges of supervised learning methods include acquiring labeled data, feature engineering and selection of classification algorithms [30], [31]. Some of the widely considered features include presence of words, word frequencies, part of speech tagging, opinion words and phrases, negations and n-grams [25], [27], [31]. For classification algorithms, support vector machines and naïve Bayes classifiers have been shown to be effective in single domain document sentiment classification [31], [32]. Despite the ability of supervised methods to adapt to the domain of the training data, these methods suffer a few shortcomings. Firstly, apart from product reviews with user-provided rating, gathering labeled training data may be costly [33]. Secondly, multiple studies have shown that classifiers trained in one domain perform poorly when applied to other domains [25, p. 40], [32, p. 2], [27], [34], [35]. When the amount of labeled training data is insufficient, unsupervised methods can be used to cluster documents based on a selected feature set [36]. Lexical-based methods use a pre-defined list of words and n-grams tagged with associated sentiment and classify documents based on the frequency or presence of sentiment words [28]. Examples of lexical-based methods include Linguistic Inquiry and Word Count (LIWC) [37], Emolex [38] and VADER [39]. Manual labeling, dictionary-based approaches and corpus-based machine learning approaches have all been used for the development of opinion lexicons. Dictionary-based approaches build a set of opinion words by finding the transitive closure of synonyms and antonyms of some initial set of words. This is commonly performed using an online lexical database, such as WordNet. Corpus-based approaches also grow a set of words from some initial set, but use a domain specific corpora and a set of linguistic constraints instead of a lexical database. As a result, the latter is better at capturing domain specific information [27]. It is worth noting that lexical-based and machine learning methods can be combined, as exemplified by Melville et al. [40]. Several sentiment analyzers using machine learning methods also incorporate lexical dictionaries. Examples of such analyzers include SentiStrength [32], [41], Stanford recursive deep model [42] and SentiWordNet [43], [44].

**Data**

**Data Collection**

Stack Exchange provides a data dump of all user contributed content, which is updated every quarter. The data dump for each website is separated into multiple zipped XML files based on the content, with each XML file representing a database table [19]. In this study, only the XML files for posts and comments were used. Table 3 lists the requirements by which data was gathered from the dump.

**Data Description**

**Questions.csv** contains the title, body, creation date, score, and owner ID for each Python question.

|  |  |
| --- | --- |
| Column Name | Description |
| Id | Unique identifier for the post. |
| Owner User Id | Only answers contain this attribute. Indicates the Id of the associated question. |
| Creation Date | Creation date of the post. |
| Score | How many upvotes are received |
| Title | The title of a question. Only questions contain this attribute. |
| Body | Content of the post. |

**Preprocessing**

Before being classified by Senti4SD, documents need to be sanitized and formatted. Text that may interfere with the classification, such as code snippets and urls, should be removed, and each document must be contained on a single line.

**General Preprocessing**

The following general preprocessing algorithm was performed on all documents.

1. Remove the bodies of all <pre>, <code> and <blockquote> tags.

2. Extract all text (excluding any HTML tags).

3. Replace every sequence of consecutive whitespace (such as spaces, tabs and line feeds) with a single space character each.

4. Remove all http and https urls.

5. Strip leading and trailing whitespace.

**Conclusions**

The results show substantial differences between the sentiment distributions of different Question types, implying a strong correlation between Question type and sentiment. However, definitive conclusions about such correlations cannot be drawn due to uncertainties in tooling errors. What can be said with some degree of certainty is that all questions are predominantly neutral. As for the programming language analysis, the results indicate that there is a correlation between programming language and sentiment. Drawing conclusions from this would however be inadvisable because of the large standard deviations of the samples and the potential subdomain specific bias of the classifier. What can be said is that there appears to be a correlation, but further research is needed for any concrete conclusions to be drawn. Above all, this study has shown that sentiment analysis of technical texts is still in its early stages. Numerous issues remain, including analyzer bias, varying preprocessing methods and possible variations of accuracy and bias between subdomains. It is possible that sentiment analysis classifiers for SE are not yet mature enough for this kind of detailed comparative study.

***References***

[1] A. Y. Chua and S. Banerjee, “Answers or no answers: Studying question answerability in stack overflow,” Journal of Information Science, vol. 41, no. 5, pp. 720–731, 2015.

[2] D. Movshovitz-Attias, Y. Movshovitz-Attias, P. Steenkiste, and C. Faloutsos, “Analysis of the reputation system and user contributions on a question answering website: Stackoverflow,” in Proceedings of the 2013 ieee/acm international conference on advances

in social networks analysis and mining, 2013, pp. 886–893.

[3] “Privileges - stack overflow,” 2018. [Online]. Available: https://stackoverflow.com/help/privileges. [Accessed: 12-Apr-2018].

[4] “Stack exchange site statistics.” [Online]. Available: https://stackexchange.com/sites.

[Accessed: 25-Mar-2018].

[5] “What are tags, and how should i use them? - stack overflow,” 2018. [Online]. Available: https://stackoverflow.com/help/tagging. [Accessed: 28-May-2018].

[6] M. De Choudhury and S. Counts, “Understanding affect in the workplace via social media,” in Proceedings of the 2013 conference on computer supported cooperative work, 2013, pp. 303–316.

[7] E. Guzman and B. Bruegge, “Towards emotional awareness in software development teams,” in Proceedings of the 2013 9th joint meeting on foundations of software engineering, 2013, pp. 671–674.

[8] J. M. George and A. P. Brief, “Feeling good-doing good: A conceptual analysis of the mood at work-organizational spontaneity relationship.” Psychological bulletin, vol. 112, no. 2, p. 310, 1992.

[9] F. Calefato, F. Lanubile, M. C. Marasciulo, and N. Novielli, “Mining successful answers in stack overflow,” in Mining software repositories (msr), 2015 ieee/acm 12th working conference on, 2015, pp. 430–433.

[10] F. Calefato, F. Lanubile, F. Maiorano, and N. Novielli, “Sentiment polarity detection for software development,” Empirical Software Engineering, pp. 1–31, 2017.

[11] F. Calefato, F. Lanubile, and N. Novielli, “How to ask for technical help? Evidencebased guidelines for writing questions on stack overflow,” Information and Software Technology, vol. 94, pp. 186–207, 2018.

[12] R. Rodricks, “Why is stack overflow so negative of late?” 2014. [Online]. Available:

https://meta.stackoverflow.com/questions/251758/why-is-stack-overflow-so-negativeof-late. [Accessed: 25-Mar-2018].

[13] K. Rudolph, “The rudeness on stack overflow is too damn high,” 2014. [Online]. Available: https://meta.stackoverflow.com/questions/262791/the-rudeness-on-stack-overflowis-too-damn-high. [Accessed: 25-Mar-2018].

[14] C. Upchurch, “Could we please be a bit nicer to new users,” 2014. [Online].

Available: https://meta.stackexchange.com/questions/9953/could-we-please-be-a-bitnicer-to-new-users. [Accessed: 25-Mar-2018].

[15] J. Hanlon, “The new new "be nice" policy ("code of conduct") - updated with your feedback,” 2014. [Online]. Available: https://meta.stackexchange.com/questions/240839/the-new-new-be-nice-policy-code-of-conduct-updated-with-your-feedback. [Accessed: 12-Apr-2018].[16] J. Hanlon, “Stack overflow isn’t very welcoming. It’s time for that to change.” 2018.[Online]. Available: https://stackoverflow.blog/2018/04/26/stack-overflow-isnt-verywelcoming-its-time-for-that-to-change/. [Accessed: 11-May-2018].[17] E. Guzman, D. Azócar, and Y. Li, “Sentiment analysis of commit comments in github: An empirical study,” in Proceedings of the 11th working conference on mining software repositories, 2014, pp. 352–355.

[18] C. Danescu-Niculescu-Mizil, M. Sudhof, D. Jurafsky, J. Leskovec, and C. Potts, “A computational approach to politeness with application to social factors,” arXiv preprintarXiv:1306.6078, 2013.

[19] “Stack exchange data dump.” [Online]. Available: https://archive.org/details/stackexchange. [Accessed: 12-Feb-2018].

[20] C. A. Thompson, “A brief introduction to natural language processing for nonlinguists,” in Learning language in logic, Springer, 2000, pp. 36–48.

[21] X. Huang, A. Acero, H.-W. Hon, and R. Reddy, Spoken language processing: A guide to theory, algorithm, and system development, vol. 95. Prentice hall PTR Upper Saddle River, 2001.

[22] P. M. Nadkarni, L. Ohno-Machado, and W. W. Chapman, “Natural language processing: An introduction,” Journal of the American Medical Informatics Association,vol. 18, no. 5, pp. 544–551, 2011.

[23] D. Yu and L. Deng, Automatic speech recognition: A deep learning approach. Springer,2015.

[24] S. Armstrong, “Corpus-based methods for nlp and translation studies,” Interpreting,vol. 2, no. 1, pp. 141–162, 1997.

[25] B. Pang, L. Lee, and others, “Opinion mining and sentiment analysis,” Foundations and Trends in Information Retrieval, vol. 2, nos. 1 – 2, pp. 1–135, 2008.

[26] B. Liu, “Sentiment analysis and opinion mining,” Synthesis lectures on human language technologies, vol. 5, no. 1, pp. 1–167, 2012.

[27] B. Liu and L. Zhang, “A survey of opinion mining and sentiment analysis,” in Mining text data, Springer, 2012, pp. 415–463.

[28] F. N. Ribeiro, M. Araújo, P. Gonçalves, M. A. Gonçalves, and F. Benevenuto,“SentiBench-a benchmark comparison of state-of-the-practice sentiment analysis methods,”EPJ Data Science, vol. 5, no. 1, p. 23, 2016.

[29] M. B. Christopher, PATTERN recognition and machine learning. Springer-Verlag New York, 2016.

[30] B. Pang, L. Lee, and S. Vaithyanathan, “Thumbs up?: Sentiment classification using machine learning techniques,” in Proceedings of the acl-02 conference on empirical methods in natural language processing-volume 10, 2002, pp. 79–86.

[31] H. Tang, S. Tan, and X. Cheng, “A survey on sentiment detection of reviews,” Expert Systems with Applications, vol. 36, no. 7, pp. 10760–10773, 2009.

[32] M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, and A. Kappas, “Sentiment strength detection in short informal text,” Journal of the Association for Information Science and Technology, vol. 61, no. 12, pp. 2544–2558, 2010.[33] P. Gonçalves, M. Araújo, F. Benevenuto, and M. Cha, “Comparing and combining sentiment analysis methods,” in Proceedings of the first acm conference on online socia networks, 2013, pp. 27–38.

[34] A. Bermingham and A. F. Smeaton, “Classifying sentiment in microblogs: Is brevity an advantage?” in Proceedings of the 19th acm international conference on information and knowledge management, 2010, pp. 1833–1836.

[35] A. Aue and M. Gamon, “Customizing sentiment classifiers to new domains: A case study,” in Proceedings of recent advances in natural language processing (ranlp), 2005,vol. 1, pp. 2–1.

[36] P. D. Turney, “Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews,” in Proceedings of the 40th annual meeting on association for computational linguistics, 2002, pp. 417–424.

[37] Y. R. Tausczik and J. W. Pennebaker, “The psychological meaning of words: LIWC and computerized text analysis methods,” Journal of language and social psychology, vol. 29, no. 1, pp. 24–54, 2010.

[38] S. M. Mohammad and P. D. Turney, “Crowdsourcing a word–emotion association lexicon,” Computational Intelligence, vol. 29, no. 3, pp. 436–465, 2013.

[39] C. H. E. Gilbert, “Vader: A parsimonious rule-based model for sentiment analysis of social media text,” in Eighth international conference on weblogs and social media (icwsm-14). Available at (20/04/16) http://comp. Social. Gatech. Edu/papers/icwsm14.

Vader. Hutto. Pdf, 2014.

[40] P. Melville, W. Gryc, and R. D. Lawrence, “Sentiment analysis of blogs by combining lexical knowledge with text classification,” in Proceedings of the 15th acm sigkdd international conference on knowledge discovery and data mining, 2009, pp. 1275–1284.

[41] M. Thelwall, K. Buckley, and G. Paltoglou, “Sentiment strength detection for thesocial web,” Journal of the Association for Information Science and Technology, vol. 63,no. 1, pp. 163–173, 2012.

[42] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Ng, and C. Potts,“Recursive deep models for semantic compositionality over a sentiment treebank,” in Proceedings of the 2013 conference on empirical methods in natural language processing,

2013, pp. 1631–1642.

[43] A. Esuli and F. Sebastiani, “SentiWordNet: A high-coverage lexical resource for opinion mining,” Evaluation, pp. 1–26, 2007.

[44] S. Baccianella, A. Esuli, and F. Sebastiani, “Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining.” in LREC, 2010, vol. 10, pp. 2200–2204.