An Exploratory Analysis of a Hybrid OSS Company's Forum in Search of Sales Leads

Myriam Munezero
Department of Computer Science
University of Helsinki
Helsinki, Finland
myriam.munezero@helsinki.fi

Tero Kojo
The Qt Company
Espoo, Finland
tero.kojo@qt.io

Tomi Männistö
Department of Computer Science
University of Helsinki
Helsinki, Finland
tomi.mannisto@helsinki.fi

Abstract*—*Background: **Online forums are instruments through which information or problems are shared and discussed, including expressions of interests and intentions.**

Objective: **In this paper, we present ongoing work aimed at analyzing the content of forum posts of a hybrid open source company that offers both free and commercial licenses, in order to help its community manager gain improved understanding of the forum discussions and sentiments and automatically discover new opportunities such as sales leads, i.e., people who are interested in buying a license. These leads can then be forwarded to the sales team for follow-up and can result in them potentially making a sale, thus increasing company revenue.**

Method: **For the analysis of the forums, an untapped channel for sales leads by the company, text analysis techniques are utilized to identify potential sales leads and the discussion topics and sentiments in those leads.**

Results: **Results of our preliminary work make a positive contribution in lessening the community manager's work in understanding the sentiment and discussion topics in the hybrid open source forum community, as well as make it easier and faster to identify potential future customers.**

Conclusion: **We believe that the results will positively contribute to improving the sales of licenses for the hybrid open source company.**

***Keywords*-Sales lead identification; text analysis; online forums; sentiment analysis; topic modeling; hybrid OSS company;**

I. Introduction

Online forums of open source software (OSS) projects play a highly important role in the whole development cycle of the projects. They contain a large amount of knowledge about problems and their solutions, as well as feedback and improvement suggestions [1]. In addition, messages among posters are exchanged or posted on a regular basis on varying topics such as requests for support, bug reports and fixes, new feature requests, etc. [2], thus making forums a rich resource that can be used to gain deeper insights and reveal potential new opportunities.

This paper presents ongoing practical work conducted with a specific hybrid open source software (OSS) company, where the company has a free and a commercial license offering. Such a hybrid structure leads to forum discussions that include inquiries, clarification questions, or expressions of interest and intentions about licenses. In particular, the hybrid OSS company has a community manager who manually peruses the forum to provide support, reply to questions, and look for ways to make improvements to the community and or software product. His job also includes looking for forum posters who express some interest in buying the commercial licenses, i.e., sales leads, and then forwarding the leads to the sales team in the company who can then follow-up with the lead and potentially make the sale. He does this not only to increase the sales of the company but to also ensure that community members have the license they need for their work. As Futrell [3] pointed out, finding sales leads or people who might be prospects is one of the most important parts of the selling process, since you can't make a sale without identifying the people to whom you will be selling. Not only that, the identification of sales leads can be beneficial for marketing, sales, and customer service business functions [4]. The manual perusal, however places a heavy burden on the community manager, and taking into account the sheer volume of posts, the task is almost impossible and costly.

Thus this research seeks to address and assist the community managers in their task of identifying sales leads. The hybrid nature of the company provides a unique context as it opens up discussions and interests around commercial licenses even among the open source communication channels, e.g., forums. In particular, the research aimed to answer the following research questions:

- RQ1: How are sales leads represented in forum posts of a hybrid OSS company?
- RQ2: How can the sales leads be automatically detected?
- RQ3: What insights are afforded to the community manager by performing the sales leads detection?

The study made use of text analysis techniques to extract meaningful information from the unstructured textual forum data, through the identification and exploration of patterns and relationships in the text [5], [6]. The main focus was on identifying sales leads in forum posts, an untapped channel for leads by the company. In addition, topic modeling was used to automatically identify discussion topics, i.e., groups of related words that approximate a real-world concept, and also perform sentiment analysis to discover the positive or negative feelings in the sales lead posts as a way to gain a broader understanding of the phenomenon and its representation.

Based on qualitative analyses performed with the company's community manager, it was observed that this work helped him gain new insights and identify opportunities, which otherwise would not have been manually possible with the huge amount of data available. For example, he was able to identify a good target group for the commercial license sales based on the discussion topics identified. Thus, this work made it easier and faster for the manager to identify potential future customers.

II. Related work

In this section, we discuss related works in sales leads identification. This includes the detection of purchase intentions, wishes, and interests. Purchase intentions were the focal point in, for instance, Gupta et al.'s [7] work. They defined *purchase intent* as a text expression showing a desire to purchase a product or service in the future. Using a binary supervised classification approach, they made use of linguistic and statistical features to classify a social post as either purchase intent or non-purchase intent.

Ramanand et al. [8] focused on detecting *buy wishes* from product reviews. The buy wishes were either suggestions made by customers or intentions to purchase a product or service. They used a rule-based approach to detect the wishes from text. Wu and He [9] also studied the problem of automatically identifying wishes in product reviews. They used a keyword strategy to first find candidate wish sentences, from which sequential patterns were then mined. The patterns were then used as features to train a classifier to automatically identify wish sentences in product reviews.

Moreover, Fu and Liu [10] presented a weakly-supervised approach to detect *consumption intent* in microblogs. They defined consumption intent as a desire or hope for something to purchase (immediate or future). They presented their detection problem as a binary classification task. To build their positive consumption intent dataset, they made use of designated hashtags (e.g., #want, #buy, #seek recommendation, etc.), and automatically collected a large number of posts and regarded them as positive instances. However, the presence of these hashtags does not necessary mean that the posts contained consumption intent.

Many of the related works as can be seen have focused more on product reviews and social media. However, not much has been done to explore sales leads in forums of hybrid OSS companies. Using similar techniques as Ramanand et al. [8], we investigated the detection of sales leads in forums of a hybrid OSS project which made for an interesting domain to be explored.

III. Research approach

The main objective of the study was to explore and analyze the content of forum posts using text analysis techniques and provide solutions for a community manager. The solutions allow for identifying the presence of any possible leads in forum posts and understanding the phenomenon in terms of how the leads are represented, the discussion topics, and sentiments present in the leads. Subsections III-A to III-C describe the techniques applied on a collection of forum posts.

A. Identifying sales leads

Having the ability to identify sales leads provides several benefits for companies such as improving the revenue stream and increasing customer base. As sales lead identification in text is a relatively unexplored area especially in forum posts of hybrid OSS companies, this work aims to bring in new insights. We particularly aim to achieve this by investigating sale lead indicators and the topics and sentiments present in those posts.

To explore the sales leads and their identification in text, we made use of a three-step approach: First, we manually collected and created a corpus of positive posts that have been agreed upon by the community manager to contain sales leads, that is, posts inquiring about buying a license or expressing interest and intention to buy a license. For example, *"I want to purchase a licence."* From this positive set, we performed qualitative analysis to extract sales lead indicators (e.g., *purchase*) which then formed as a seed list of lead indicators. The seed list was then expanded manually with a sample of synonyms from the Merriam-Webster thesaurus[1].

The second step involved using the seed list to retrieve posts that matched one or more items in the seed list. We evaluated the retrieved posts and measured the precision of the resulting posts. Those posts that were retrieved as positive and were found to be positive by the researcher, were added to the positive sales leads corpus. In future, as the corpus becomes large enough, we plan to make use of machine learning algorithms to automatically detect those posts that contain sales leads in the forums.

B. Extraction of forum topics

Our approach makes use of topic modeling, a suite of algorithms that automatically find the overreaching topics from a collection of textual content, without the need for tags, training data, or predefined taxonomies [11], [12].

Identifying topics in a collection of text helps identify characteristics of the communication which can help management better understand the needs of the forum members. By being privy to this, positive interventions can be made where necessary. For example, Barua et al. [13] made use of topic modeling to automatically discover the main topics in developer discussions in StackOverflow. In their work, they identified that the developer discussion topics range widely from jobs to version control systems to code syntax.

Specifically for our work, we used a popular topic modeling technique called latent Dirichlet allocation (LDA), a statistical topic modeling technique, to automatically discover the main topics present in the forum posts (i.e., groups of related words that approximate a real-world concept). LDA represents topics as probability distributions over the words in the corpus, and it represents documents as probability distributions over the

[1] https://www.merriam-webster.com/thesaurus

discovered topics. LDA creates topics when it finds sets of words that tend to co-occur frequently in the documents of the corpus. Often, the words in a discovered topic are semantically related, which gives meaning to the topic as a whole [13]. Thus we use the discovered topics, as approximations of the discussion topics in forum posts.

C. Extracting sentiments

Sentiment analysis (SA) is framed within the area of natural language processing (NLP) and is broadly defined by Pang and Lee [14] as the computational treatment of opinions, feelings, emotions, and subjectivity in texts. Over the past years, a large number of SA programs have been developed to discover the sentiment content of texts in various genres including movie reviews [15], student diaries [16], education forums [17], [18], and developer platforms. For example, in Parastou et al. [19], they extracted the sentiment from user and developer mailing lists of two of the most successful and mature projects of the Apache software foundation. Their results showed that user and developer mailing lists carry both positive and negative sentiment and have a slightly different focus.

Generally, two main methods exist for the analysis of emotions within the NLP community: word lists-based and machine learning-based SA. In this work, we extract sentiments using a naive Bayes machine learning classifier that takes a piece of text as input and classifies it as either positive or negative. The classifier has been trained on collection of phrases that have been categorized according to their contextual polarity (i.e., positive and negative) [20].

IV. Results

We evaluated our approach detailed in Section III on a database dump of forum posts. Section IV-A details the dataset used for the evaluation and the results are presented in Section IV-B.

A. Dataset

For the evaluation of our approach, we made use of forum data from a hybrid OSS company, The Qt Company[2]. The company's product, Qt software framework, is an open source software that is used to create platform-independent applications for Android, iOS, and Windows operating system environments. Its users include independent application developers and companies from a wide variety of industries including electronics, automotive, defense, and media. The Qt Company has a hybrid licensing model which offers both Open Source and commercial options. The commercial licenses allow for making applications proprietary, to access new software components, and to receive varying levels of support from the company. Development of Qt software is driven by a versatile, open community for which the company provides software development tools.

The Qt Company forum consists of six channels. For the analysis in this work, we particularly made use of a data dump from the most active forum channel, the *Qt development* channel which is also the channel the community manager mostly peruses. The discussions in there include everything on development, from desktop, mobile, cloud, games, tools, 3rd party libraries, etc. The data dump, which was in BSON format was first transformed to JSON format and exported into CSV format for analysis. Only the main post text was of interest, other data contained in the data dump such as ids, post number were removed as they were not the focus of the current study, but will be taken into consideration in future developments, as well as the other forum channels. Moreover, as anyone on the Qt forums is registered with a Qt Company account, it makes the identified leads and forwarded leads to the sales team more trustworthy. In total, the CSV file contained 612,135 posts.

[2]https://www.qt.io/

TABLE I
SALES LEAD INDICATORS

Can I BUY	Can I SUBSCRIBE
Do I need to BUY	I have to BUY
How can I BUY	Can I PAY for
I want to PURCHASE	Thinking of PURCHASING
If I BUY	How to ACQUIRE
Are there any DISCOUNTS	Is there any available DISCOUNT

B. Identifying sales leads

1) Sales lead indicators: A lead in this work is considered as a potential buyer. In this paper, we focus on identifying explicit leads so that there is less subjectivity involved in the interpretation. For example, in the two example posts below the author clearly inquires or gives an expression of interest in buying a license.

Ex1: "Another little question: can I buy 'startups license' even if I'm a private or only a company can?" (Inquiry about buying license)

Ex2: "I want to purchase Qt license, should I wait for new version like something 5.0....? or 4.7 license will work for later versions too. If anybody have some idea of pricing for multiple platforms." (Interest in purchasing license)

Based on our first step in our research approach, we manually analyzed 300 forum posts. An initial set of thirteen posts that could positively be identified as positive for sales leads were manually collected and analyzed for indicators. In total, we identified 12 indicators which are summarized in Table I.

As can be seen in Table I, the indicators are expressed in three to five word phrases with the keywords and phrases being very similar. The keywords are capitalized and in bold in Table I. Each indicator in Table I was in regards to the self subject ('I') and object commercial license (e.g., can I buy [license]).

In the second step, we made use of the six keywords in Table I (i.e., buy, subscribe, pay, purchase, acquire, and discount), and searched for relevant synonyms from the Merriam-Webster dictionary to expand the list of keywords. Note that we took purchase and purchasing as one keyword. Similarly with discounts and discount. Based on our search, we found and included the terms 'get' and 'obtain'. Thus, in total

we ended up with eight keywords as part of the sales lead indicators. The set of sales lead keywords is small, indicating that the manner of expressing interest or intention is often similar.

Using this list of indicators, a python script was written to retrieve all posts that contained at least one of the terms from the list of indicators. In total, 206 posts were retrieved. Considering our dataset of about 600,000 posts, it is clear that the number of potential sales leads is very small. However, considering that The Qt Company has a dedicated sales channel where those interested in the commercial licenses can directly contact the sales team, it does explain the low number of sales leads identified in the forum posts, as the forum is not the primary channel for showing interest in buying the commercial licenses. Leads from this untapped channel would, however, contribute to the company sales.

As this was a relatively small set, manual analysis was performed to identify whether the retrieved posts were actually positive. A set of only 51 posts were identified as being positive, which means that of the total retrieved posts, 25% of them were positive based on our search. To form the final dataset of positive sales leads, the identified 51 posts together with our initial set of 13 posts were combined. Three posts had to be removed because of duplication. Thus the total positive dataset had 61 posts.

Moreover, to explore the content of this positive dataset, we performed further analyses. Results of these are presented in the next subsections.

2) Sales leads dataset description: The positive sales lead posts were first pre-processed using R, the statistical computing language and environment, to prepare it for the analysis. Punctuation, digits, URL links, unnecessary spaces and specific characters were first removed. Then the posts were put in the lower case, and each word was stemmed using Porter's [21] stemming algorithm. English stop words were also removed from the dataset. Moreover, as software-related conversations often contain large passages of source code snippets, configuration files, automatically created debug output etc., we removed the code snippets as they contain similar programming language syntax and keywords, and these do not help topic models to find useful topics nor sentiment [22].

Using R to explore the positive dataset, we identified that the 51 posts consisted of 1,301 words. Figure 1 illustrates the top 100 most frequent unique words.

As can be seen in figure 1, the most frequent words include 'can', 'app', 'commercial', 'buy', 'use', and 'will'. Upon further analysis of the dataset, the term 'can' for instance was most used as a means of asking questions, where the poster was wondering whether they 'can' do something or perform some action, or 'use' some method or approach, while 'app' showed the focus of the posts, i.e., asking questions or sharing information about developing apps.

Thus, the frequent words gave some context of the conversation around the sales leads. Furthermore, in comparison to our identified sales lead indicators, the most frequent terms were 'buy' and 'purchase'.

Fig. 1. Wordcloud displaying the top frequent 100 words.

3) Sentiments: Using the positive list, we further explored the sentiments using the naive Bayes classifier described in Section III-C. The unit of analysis was at the post level. Figure 2 shows the distribution of posts with each sentiment category.

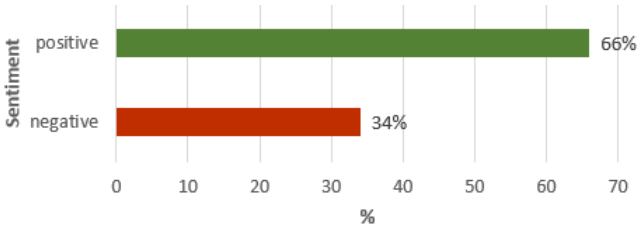

Fig. 2. Sentiments in the sales lead posts.

In Figure 2, we found that the posts contained more positive sentiments than negative sentiments. We identified that there were more positive posts, which is perhaps intuitive in that when people show or express interest in buying a product, they are more positive towards that product than negative.

4) Topics: Using the positive list to explore and identify discussion topics in forums, we made use of the LDA function with Gibbs sampling from R's topic models package, which includes setting a value for the number of topics to output. Unfortunately, there is no value for the number of topics that is appropriate in all tasks and datasets [23], [24]. A large value for the number of topics will result in fine-grained and more detailed topics while a smaller number will produce coarser-grained, more general topics. In this work, we aim for medium granularity, so that the topics capture the broad trends in our dataset while remaining distinct from each other. We experimented with a number of different values for k including 5, 10, 15, and 20. Based on inspecting the results, we decided to settle on 10 as it gave us sufficient granularity to reason about and identify the discussion topics in the posts.

For clarity in presenting the results, we show the first five topics discovered by our topic model in Table II along with a

sample of five words belonging to each topic.

As can be seen in Table II, LDA outputs for each topic, a set of words sorted in terms of their likelihood of belonging to that topic. As LDA does not generate meaningful labels for each topic, a manual study of the set of words in the context of the forum posts was performed. This revealed that, for instance: Topic 1 is related to inquiries or requirements of support, installing and running programs, and iOS-related discussions, Topic 2 is related to monetary concerns and aspects (pay, pricing, subscription, cost), while Topic 3 is on

TABLE II
RESULTS OF THE LDA

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
ios	code	project	can	users
support	pay	charts	app	gui
program	pricing	desktop	commercial	release
install	subscription	linux	buy	software
run	cost	experience	use	open

certain technologies and projects.

V. DISCUSSION AND THE MANAGER'S PERSPECTIVE

In this paper, we have presented preliminary results of ongoing work aimed at identifying the presence of sales leads in forum posts of a hybrid OSS company. This work makes a valuable contribution to the company, as the forum was an untapped channel for sales leads by the company. Based on the results, we elicited an expert opinion from the community manager about our empirical results presented in Section IV. An interview was conducted with the Qt's community manager where he was provided with the results of the leads, sentiments, and discussion topics and was asked to assess and discuss the impact and value of the results.

A. Potential for growth – sales leads

In our analysis, we positively identified that there are possible sales leads present in the posts. Following our RQ1, we identified the sales lead indicators and how they are represented in the forum posts. We found that they are expressed over three to five word phrases in the whole post and mostly include one of the eight identified keywords such as 'buy', 'purchase', 'obtain', etc., with 'buy' being the most frequent keyword. Based on our RQ2, the eight keywords can be used by the community manager to search and retrieve those posts that could be potential sales leads, which as we saw in our experiments resulted in a more manageable dataset for manual perusal.

By detecting these posts, following from RQ3, the community manager can then easily send them to the sales team, which if followed through, could lead to the acquisition of new customers and thus lead to improved product sales. As the manager says, "sending the lead is not a big effort, the hard part is really finding the possible leads." Furthermore, since anyone on the Qt forum is registered with a Qt account, the sales leads are more trustworthy. Thus this work has made a positive contribution to lessen the community manager's work in hybrid open source communities, by making it easier and faster to identify potential future customers, and in turn get additional support for the project in the future.

In this work, we only looked at explicit linguistic indicators of interest and intentions to buy a license, however, there might be other implicit indications which we plan to investigate further. For example, a poster might ask a question regarding a task they want to perform or an app they would like to build, and from the questions or specifications, it would be possible to see whether they would benefit more from a commercial license, thus becoming a potential sales lead. But this was beyond the scope of this work.

B. Sentiments

In addition to exploring the sales lead indicators and keywords, we identified the sentiments present in the sales lead posts and found that the leads are represented more positively than negatively. Being able to identify sentiments, as the manager pointed out, can inform decision making in terms of how to react to posts. Normally, the manager used their own heuristics to gauge the sentiment in the forum community, however, with our work, they can now have empirical evidence of the distribution of sentiments which can be reported and shared with company members. In future, the sentiment data can be used to compare changes over time or before and after a version release.

C. Topics

The manager found the topics identified in the forums to be clear. In particular, based on the wordcloud and the discussion topics, the community manager stated that he was able to gain new insights, for example, he was able to identify based on the discussion topic 4 (see Table II), that mobile developers would be a good target market for commercial licenses. This as he stated "would increase [the] likelihood of pointing sales to possible customers." In addition, he particularly found the frequent words illustrated as a wordcloud to be important in bringing out the relative importance of words. "It makes it easier to see what to take on first," that is, get at the most important things first. Future work involves analyzing the topics to see if they could be indicators of other phenomena.

Thus, based on feedback from the company's community manager, the solutions presented in this work helped him gain new insights and identify opportunities, which otherwise would not have been manually possible with the amount of data available. We acknowledge that the external validity of the empirical study is comprised by the single-case study design and involvement of only one community manager. Even though generalizability was not the main goal of the study, since hybrid OSS projects typically form around business-driven, platform-like products, the ecosystem can be considered representative.Thus, our approach and results may be useful for other companies with similar characteristics.

In addition, only a few hundred posts were investigated at this time when looking for lead indicators, thus potentially

missing some indicators. This is an area where future work will also focus. However, using synonyms to expand the identified indicators was an effort to broaden the scope.

VI. Conclusion

The paper explored the use of text analysis techniques to support the community manager of a hybrid OSS company in their daily work, which includes looking for potential customers. The results make a positive contribution to lessen the community manager's work in understanding the sentiment and discussion topics in the hybrid open source forum community, as well as make it easier and faster to identify potential future customers, thus increasing company revenue.

Based on feedback from the company's community manager, the results presented in this work helped him gain new insights and identify opportunities, which otherwise would not have been manually possible with the amount of data available.

Future work involves integrating the solutions into the forum posts such that each incoming post is analyzed and labeled as a sales lead or not, being of negative, positive, or neutral sentiment, and with discussion topics in the post made visible. Furthermore, we will explore the use of the sale lead corpus and the indicators to increase it to a sizeable corpus for training machine learning algorithms. In this direction, we realize that practical challenges such as having a big imbalanced dataset would need to be addressed.

For more in-depth analysis of the analysis, we also plan to investigate other information that may be indicators for sales leads, other than just the textual post or those that might increase the certainty of the identified lead. For example, personal information about the poster, their activities on other company channels, and frequency of posting.

Acknowledgment

This study was supported by the Need for Speed research program of DIMECC funded by Tekes as well as the European Union's Horizon 2020 research and innovation programme under grant agreement No 732463.

References

[1] B. Almeida, S. Ananiadou, A. Bagnato, A. B. Barbero, J. Di Rocco, D. Di Ruscio, D. S. Kolovos, I. Korkontzelos, S. Hansen, P. Maló *et al.*, "Ossmeter: Automated measurement and analysis of open source software." in *STAF Projects Showcase*, 2015, pp. 36–43.

[2] F. Ahmed, P. Campbell, A. Jaffar, and L. F. Capretz, "Myths and realities about online forums in open source software development: an empirical study," *The Open Software Engineering Journal*, vol. 4, pp. 52–63, 2010.

[3] C. Futrell, *ABCs of Relationship Selling*. McGraw-Hill Higher Education, 2005.

[4] C. S. Carlos and M. Yalamanchi, "Intention analysis for sales, marketing and customer service." in *COLING (Demos)*, 2012, pp. 33–40.

[5] W. He, S. Zha, and L. Li, "Social media competitive analysis and text mining: A case study in the pizza industry," *International Journal of Information Management*, vol. 33, no. 3, pp. 464–472, 2013.

[6] W. Song and S. C. Park, "A novel document clustering model based on latent semantic analysis," in *Semantics, Knowledge and Grid, Third International Conference on*. IEEE, 2007, pp. 539–542.

[7] V. Gupta, D. Varshney, H. Jhamtani, D. Kedia, and S. Karwa, "Identifying purchase intent from social posts." in *ICWSM*, 2014, pp. 180–186.

[8] J. Ramanand, K. Bhavsar, and N. Pedanekar, "Wishful thinking: finding suggestions and'buy'wishes from product reviews," in *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, 2010, pp. 54–61.

[9] X. Wu and Z. He, "Identifying wish sentence in product reviews," *Journal of Computational Information Systems*, vol. 7, no. 5, pp. 1607–1613, 2011.

[10] B. Fu and T. Liu, "Weakly-supervised consumption intent detection in microblogs," *Journal of Computational Information Systems*, vol. 6, no. 9, pp. 2423–2431, 2013.

[11] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proceedings of the National academy of Sciences*, vol. 101, no. suppl 1, pp. 5228–5235, 2004.

[12] D. M. Blei and J. Lafferty, *Topic models. Text mining: theory and applications*. Taylor and Francis London, 2009.

[13] A. Barua, S. W. Thomas, and A. E. Hassan, "What are developers talking about? an analysis of topics and trends in stack overflow," *Empirical Software Engineering*, vol. 19, no. 3, pp. 619–654, 2014.

[14] B. Pang, L. Lee *et al.*, "Opinion mining and sentiment analysis," *Foundations and Trends® in Information Retrieval*, vol. 2, no. 1–2, pp. 1–135, 2008.

[15] B. Pang and L. Lee, "A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts," in *Proceedings of the 42nd annual meeting on Association for Computational Linguistics*, 2004, p. 271.

[16] M. Munezero, C. S. Montero, M. Mozgovoy, and E. Sutinen, "Exploiting sentiment analysis to track emotions in students' learning diaries," in *Proceedings of the 13th Koli Calling International Conference on Computing Education Research*. ACM, 2013, pp. 145–152.

[17] T. Zarra, R. Chiheb, R. Faizi, and A. El Afia, "Using textual similarity and sentiment analysis in discussions forums to enhance learning," *International Journal of Software Engineering and Its Applications*, vol. 10, no. 1, pp. 191–200, 2016.

[18] A. Ezen-Can, K. E. Boyer, S. Kellogg, and S. Booth, "Unsupervised modeling for understanding mooc discussion forums: a learning analytics approach," in *Proceedings of the fifth international conference on learning analytics and knowledge*. ACM, 2015, pp. 146–150.

[19] P. Tourani, Y. Jiang, and B. Adams, "Monitoring sentiment in open source mailing lists: exploratory study on the apache ecosystem," in *Proceedings of 24th annual international conference on computer science and software engineering*. IBM Corp., 2014, pp. 34–44.

[20] E. Riloff and J. Wiebe, "Learning extraction patterns for subjective expressions," in *Proceedings of the 2003 conference on Empirical methods in natural language processing*. Association for Computational Linguistics, 2003, pp. 105–112.

[21] M. F. Porter, "An algorithm for suffix stripping," *Program*, vol. 14, no. 3, pp. 130–137, 1980.

[22] S. W. Thomas, "Mining software repositories using topic models," in *Proceedings of the 33rd International Conference on Software Engineering*. ACM, 2011, pp. 1138–1139.

[23] H. M. Wallach, I. Murray, R. Salakhutdinov, and D. Mimno, "Evaluation methods for topic models," in *Proceedings of the 26th annual international conference on machine learning*. ACM, 2009, pp. 1105–1112.

[24] S. Grant and J. R. Cordy, "Estimating the optimal number of latent concepts in source code analysis," in *Source Code Analysis and Manipulation (SCAM), 2010 10th IEEE Working Conference on*. IEEE, 2010, pp. 65–74.