

感情分析による OSS プロジェクト中断の予測に向けた調査

山下 一寛^{1,a)} 亀井 靖高^{1,b)} 鵜林 尚靖^{1,c)}

概要: オープンソースソフトウェア (OSS) は、エンドユーザのみならず、企業でも利用され、非常に重要な役割を果たしている。しかし、OSS の開発者は自由に開発を辞めることができるため、OSS の開発・保守がいつまで続くかわからないことが利用者にとって大きな問題である。そのため、OSS プロジェクトの継続性を計測する研究が行われているが、主にプロジェクトの活動量を用いて計測をしており、開発者の興味に着目した研究は無い。そこで本研究では、感情分析を行うことにより、開発者の興味の推移を測ることを試みる。本稿では特に、プロジェクトリーダーの離脱により中断したプロジェクトを対象に、プロジェクトリーダーのコミットコメントを感情分析し、推移を計測する。その結果、コミットコメントは時系列に関わらず、ニュートラルな傾向を示し、プロジェクトの中断や継続との関連は無いことがわかった。

キーワード: 感情分析, オープンソースソフトウェア

Toward Predicting Termination of Open Source Software Project Using Sentiment Analysis

KAZUHIRO YAMASHITA^{1,a)} YASUTAKA KAMEI^{1,b)} NAOYASU UBAYASHI^{1,c)}

Abstract: Open Source Software (OSS) plays a significant role in our daily life, since it is widely used for not only end user's applications but also business's infrastructure. However, there are still some problems that users face when using OSS systems. For example, users do not know when OSS projects stop the development and maintenance activities because that OSS developers are not employee of the project and can quit the project freely. Previous studies work on survivability of OSS projects and mainly focus on the project activities. To the best of our knowledge, there is no research focusing on measuring interests of developers to projects. In this paper, we try to measure transitions of interests using sentiment analysis. For the first step, we only target commit comments of project leaders whose projects were abandoned due to his leave. From the analysis, we find that most of commit comments have neutral sentiment score and it does not depend on time.

Keywords: Sentiment Analysis, Open Source Software

1. はじめに

オープンソースソフトウェア (OSS) は、エンドユーザのみならず、企業でも利用され、非常に重要な役割を果たしている。Black Duck による調査によると、78%の企業が OSS システムを利用しており^{*1}、また、情報処理推進機構

の調査によると、国内で 66.8%の企業が OSS システムを利用している^{*2}。

OSS はコミュニティによって開発・保守が行われるため、企業にとっては同等のソフトウェアを開発・保守するためのコストを削減をすることができるといったメリットがある。しかし、OSS コミュニティに参加する多くの開発者は、仕事としてではなく、自身の興味に従いそのコミュ

¹ 九州大学 大学院システム情報科学府

^{a)} yamashita@posl.ait.kyushu-u.ac.jp

^{b)} kamei@ait.kyushu-u.ac.jp

^{c)} ubayashi@ait.kyushu-u.ac.jp

^{*1} <https://www.blackducksoftware.com/>

future-of-open-source

^{*2} 情報処理推進機構, “第 3 回オープンソースソフトウェア活用ビジネス実態調査 調査報告書”, 2010

ニティに参加しているため、いつでもコミュニティへの貢献を辞めることができる。そのため、企業がOSSを利用する上での大きな懸念の1つとして、利用するOSSプロジェクトがいつまで継続し、開発・保守を続けるかわからないということがあげられる^{*2}。

この問題に対応するため、プロジェクトの継続性に関する研究 [3][4][17] が行われている。例えば、Raja と Tretter [17] は、*Viability Index* というリリース数や修正時間やプロジェクトの組織構造などからプロジェクトの継続性を算出する指標を提案した。しかし、既存研究はプロジェクトの活動から継続性を判断する研究が多く、開発者がプロジェクトに対する興味を失い去っていく場合は予測できない。

そこで、本研究では開発者の感情の変化に着目し、プロジェクトに対する興味の指標として適用できるかを調査する。感情の計測には感情分析 (Sentiment Analysis) を用いる。人はコミュニケーションを取る際に、対象に対してどのように感じているかの感情が単語に自動的に含まれ [12]、感情は口頭でのコミュニケーションのみならず、コンピュータを用いたコミュニケーションにも表れる [20]。そのため、プロジェクトに対して興味が強いうちはポジティブな単語を使い、興味がなくなるとネガティブな単語を使うといったような変化が表れ、開発者のプロジェクトに対する興味を計測できるのではないかと考えた。

本稿では、開発において最も基本的な活動であるソフトウェアに対する変更 (コミット) に関するコメントを対象とする。Guzman ら [7] によって、コミットコメントは全体的にニュートラルな傾向を持つことが指摘されている。しかし、本稿では、プロジェクトリーダー^{*3} が辞めたことにより中断したプロジェクトを対象とし、プロジェクトリーダーのコミットコメントを時系列で分析をすることにより、感情がどのように変化するかにも着目する。また、比較のために継続しているプロジェクトのプロジェクトリーダーのコミットコメントも対象とする。

以下、2章では関連研究について紹介する。3章では、本稿で用いたデータセットについて説明し、4章では、本稿で行ったケーススタディの設定とその結果について説明する。5章で結果について議論を行い、6章で本稿の妥当性への脅威について説明する。そして、最後に7章で本稿をまとめる。

2. 関連研究

本章では、本稿と関連する研究について紹介する。

2.1 感情分析

感情分析はソフトウェア工学の研究にも多く用いられて

^{*3} 本稿ではリポジトリのオーナーをプロジェクトリーダーとして扱う。

いる。Guzman ら [7][8] は、GitHub で公開されているプロジェクトのコミットコメントを感情分析し、言語間や日付・時間帯などと感情との関連について調べた。その結果、コミットコメントは全体的にニュートラルな傾向を持つことや、スターの数とポジティブなコメントは弱い相関を持つことを明らかにした。

Tourani と Adams [22] は、レビュー中の議論や問題に対する議論に関する指標の1つとして、感情分析を用い、議論に関する指標を取り入れることによって、バグ予測モデルの精度が向上することを確認した。また、OSS プロジェクトのメーリングリストに対して感情分析を適用した [23]。メーリングリストに対して感情分析を適用することにより、6種類のポジティブなメールのカテゴリ、4種類のネガティブなカテゴリに分類できることを見つけた。

Ortu ら [15] は、人々の気持ちや礼儀正しさがバグや不具合などの問題の修正時間に影響するかを調査し、極端にポジティブやネガティブなコメントを含む問題は短時間で修正されることを見つけた。

感情分析を行うためのツールも複数存在する。本稿では、感情分析のために、SentiStrength というツールを用いたが、Python のNLTK ライブラリ^{*4} や OpinionFinder [25] を用いることもできる。Pletea ら [16] はNLTK ライブラリを用い、GitHub 上のセキュリティに関する議論について分析を行った。また、Positive/Negative に分類するだけでなく、Google-Profile of Mood States (GPOMS) を用いることにより、6種類の感情因子 (Calm, Alert, Sure, Vital, Kind, Happy) に分類することができる [1]。

本稿では、プロジェクトリーダーが離脱したことにより、中断したプロジェクトに着目し、プロジェクトリーダーの感情の推移を計測し、プロジェクト中断の推測に利用できないかを考察する。

2.2 プロジェクトの継続性

Chengalur-Smith ら [3][4] は、2,772 プロジェクトを用いた実証的研究から開発者の興味を惹きつける事がプロジェクトの継続性に影響していることを発見した。また、Raja と Tretter [17] は、*Viability Index* (VI) と呼ぶプロジェクトの継続能力を計測するためのメトリクスを提案した。このメトリクスは、*Vigor* (進化する能力)、*Resilience* (問題に対処する能力)、*Organization* (プロジェクトの構造) の3つの要素からなっており、 $VI = -3.834 + 0.36Vigor + 0.232Resil + 0.5Org$ という式でVIは計算される。

OSS プロジェクトはコミュニティによって開発・保守されることから、コミュニティの継続がプロジェクトの継続に影響するとし、コミュニティに着目した研究も存在する。

^{*4} <http://www.nltk.org/>

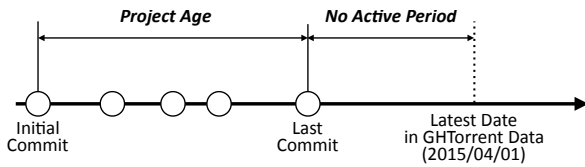


図 1 プロジェクトの期間に関する指標の概要

その一例として、良いプロジェクト管理方法を実践することによって、プロジェクトが継続していくという推測が得られている [2][11].

本稿では、プロジェクトの活動ではなく、開発者（特にプロジェクトリーダー）の感情に着目し、その変化からプロジェクト中断を推測できないかを考察する。

3. データセット

本章では、本稿で用いたデータセットについて説明する。本稿では、Gousios らによって提供されている GHTorrent [6] のデータを用いて実験を進める。GHTorrent のデータは、2015 年 4 月 1 日に作成されたものを利用した（つまり、2015 年 4 月 2 日以降のデータは含まれない）。このデータセットには、合計 12,889,617 リポジトリ分のデータが含まれている。

3.1 プロジェクト選択

(1) Fork リポジトリの除去

GitHub では、2 種類のリポジトリ（つまり、*main* リポジトリと *fork* リポジトリ）が存在する。Fork リポジトリは、*main* リポジトリをコピーしたものであり、*main* リポジトリで進んでいる開発に影響を与えること無く自由に変更を行うことができる^{*5}。GitHub では、pull request を送ることにより、*main* リポジトリの開発者の合意によって、fork リポジトリで行った変更を *main* リポジトリに反映させることができる。合意を得られた pull request は、*main* リポジトリに反映されるため、*main* リポジトリのみを実験の対象として考慮すれば良い。

Fork リポジトリを除くと、6,603,923 リポジトリがデータセットに残った。

(2) ソフトウェア開発ではないリポジトリの除去

リポジトリが *main* リポジトリであるとしても、そのリポジトリがソフトウェア開発として利用されているリポジトリとは限らない [9]。例えば、SourceForge などの GitHub 以外のホスティングサービスで公開されているリポジトリを GitHub にそのまま反映しているリポジトリや設定ファイルなどを保存しているだけのリポジトリが存在する。そこで、本稿では表 1 に示す正規表現、キーワードを用いてソフトウェア開発を行っていないプロジェクトを除去した。

その結果、3,600,532 リポジトリが残った。

(3) 新しいプロジェクトの除去

^{*5} <https://help.github.com/articles/fork-a-repo/>

表 1 ソフトウェア開発では無いプロジェクトを発見するために用いた正規表現・キーワード

カテゴリ	正規表現 [9]	対象
Mirror Of	<code>mirror of .* repo git repo of</code>	Description
Sourceforge	<code>sourceforge sf .net</code>	
Bitbucket	<code>bitbucket</code>	
Subversion	<code>\W(svn subversion)\W</code>	
Mercurial	<code>\W(mercurial hg)\W</code>	
CVS	<code>\Wcvs\W</code>	
カテゴリ	キーワード	対象
Dotfiles	dotfile, config, vimrc...	Project Name
Web	homepage, website, github.io	
Sample	sample, demo, template, example	
Backup	backup	
Tutorial	tutorial, test	

今回、我々は十分な期間での感情の変化を分析するため、3 年以上開発が継続しているプロジェクトを長期間続いているプロジェクトとし、3 年よりも短いプロジェクトを研究対象から除いた。プロジェクトの継続期間の指標として本稿では、最初のコミットから最後のコミットまでの期間を用いた（図 1: Project Age）。

継続期間が 3 年未満のプロジェクトを除いた結果、66,257 リポジトリが残った。

(4) ユーザからのフィードバックが少ないプロジェクトの除去

GitHub は、GitHub 上で効率的に共同作業するための機能が 2 つある。1 つ目は pull request である。Fork リポジトリの開発者は、pull request を送ることにより、プロジェクトに参加している他の開発者に対して自身が行った変更を *main* リポジトリに適用する意思があることを知らせる事ができる。そして、*main* リポジトリの開発者は、その変更をレビューし、議論を行い、場合によっては追加の変更を加える事ができる^{*6}。2 つ目は issue である。ユーザは GitHub 上でソフトウェアの問題点の報告やその問題点の解決を開発者に依頼することができる^{*7}。

Pull request と issue の機能はユーザのフィードバックを取り入れることに用いられる。そのため、一定数以上の pull request や issue を持つプロジェクトは多くのユーザに利用され、中断した際により大きな影響をユーザに与える。本稿では、pull request と issue の数が両方とも 100 以下であるプロジェクトを除去した。最終的に、3,722 プロジェクトがデータセットに残った。

4. ケーススタディ

4.1 アプローチ

中断したプロジェクト：本稿では、中断したプロジェクト

^{*6} <https://help.github.com/articles/using-pull-requests/>

^{*7} <https://help.github.com/articles/about-issues/>

表 2 プロジェクト中断の理由

カテゴリ	理由	根拠	プロジェクト数
Survived	Repository Moved	“Repository moved to https://github.com/refinery/refinerycms ” (refinerycms プロジェクト)	20
	Keep Working	2015年4月1日以降にコミット有り	21
Terminated	Leader Quit	“I am no longer working on this library.” (asi-http-request プロジェクト)	5
	Dependency Problem	“The msn messenger servers have been stopped (...) emesene msn backend will stop working soon” (emesene プロジェクト)	2
	Became Closed Source Project	“TotalTerminal is no longer open-source” (totalterminal プロジェクト)	1
	Canceled	“UseBB 2 (...) but got cancelled in October 2012.” (usebb プロジェクト)	1
	Unknown	-	44
Total	-	-	94

表 3 プロジェクト情報

カテゴリ	プロジェクト名	LOC	開発者数	プログラミング言語	Stars
Terminated	asi-http-request	15,455	65	Objective-C	5,680
	AppSales-Mobile	19,308	52	Objective-C	1,916
	knockout.mapping	5,352	25	JavaScript	509
	machinist	692	24	Ruby	1,134
	4chan-x	9,770	15	JavaScript	62
Survived	typhoeus	4,051	124	Ruby	2,787
	powder	498	45	Ruby	1,223
	jsonapi	16,476	29	Java	209
	iCarousel	24,616	17	Objective-C	7,336
	vagrant	53,481	798	Ruby	12,073

のプロジェクトリーダーのコミットを対象に感情分析を行い、プロジェクト開始当初と中断間際の感情の違いを比較し、中断が感情分析から予測できるかを考察する。そこで、3章で説明したフィルタリングの結果残った3,722プロジェクトの中から中断しているプロジェクトを取り出す。

中断しているか否かの判断には、Samoladasら [19] や Evangelopoulosら [5] と同様にコミットの無い期間を用いる。我々は先行研究 [26] において、GitHub上に公開されているプロジェクトの新規バージョンのリリースまでの期間について調査を行った。その結果、メジャーリリースの期間は、69日(第1四分位数)から316日(第3四分位数)であった。リリースの期間の幅が非常に広いので、同様にコミットの期間も幅が広いと考え、図1に示す通り、最後のコミットからGHTorrentデータのダンプされた日である2015年4月1日までを活動の無い期間とし、その期間が1年以上続いている場合、そのプロジェクトは中断しているとした。

その後、中断しているプロジェクトが中断した理由を、GitHubページ、公式ページ、Issues等を調査することにより確認する。そして、中断した理由がプロジェクトリーダーの離脱であるプロジェクトのコミットコメントに対して感情分析を行う。本稿では、リポジトリのオーナーをプロジェクトリーダーとしている。また、リポジトリのオーナー

がチーム*8である場合、プロジェクトリーダーが明示されている場合はその表示に従い、明示されていない場合はチームメンバーのうち最もコミット数の多いメンバーをプロジェクトリーダーとする。例えば、“typhoeus/typhoeus”プロジェクトはtyphoeusチームがオーナーであるが、ライセンスのコピーライト表示を参考にプロジェクトリーダーを決定した。また、比較のため、継続しているプロジェクトからランダムに同数のプロジェクトを選び、同様に感情分析を行う。

感情分析：感情分析には、SentiStrength*9を用いる。SentiStrengthは、短い、綺麗ではない文章を扱うことに長けており、Twitterや映画のレビューなどの短い文章でその文章を書いた人の感情を正確に示した [21]。そのため、コミットコメントという短い文章を扱うのに適していると考えた。

SentiStrengthは、負の感情を表す単語には[-5, -1]の間の値を、正の感情を表す単語には[1, 5]の値を付与する。SentiStrengthは、初期値では正と負それぞれの強さの数値を出す。本稿では、“single scale”オプションを使い1つにまとめた値を出力し、その値を利用する。例えば、“I love you but hate the current political climate.”という文章の場合、SentiStrengthで解析を行うと、“I love[3] you

*8 <https://help.github.com/articles/creating-a-team/>

*9 <http://sentistrength.wlv.ac.uk/>

but hate[-4] the current political climate.”という様に、感情に関連する単語にその感情の強さを表す値が付く。デフォルトの設定では、出力として [3, -4] という様に正、負のそれぞれの値が出力されるが、“single scale” オプションを利用した場合、[-1] という文章全体の感情を表した値が出力される。

プロジェクトリーダーのコミットのコメントに対して、SentiStrength を適用し、各コミットの感情の値を計測した後、コミットを時系列に従い、ソートし、感情の変化を観察する。また、プロジェクトの前半と後半とで感情に変化があるかを確かめるため、プロジェクトの期間を前半と後半の2分割にし^{*10}、それぞれのコミットの感情の値を比較した。

4.2 実験結果

中断したプロジェクト :3,722 プロジェクトのうち、1年以上コミットが無いプロジェクトは94プロジェクトであった。94プロジェクトを調査した結果判明した、プロジェクト中断の理由を表2に示す。今回の調査では、明示的に理由が示されているプロジェクトのみをカテゴリ分けし、理由が見つけれなかったプロジェクトは“Unknown”カテゴリとした。

調査の結果、表2の上部に示す通り、41プロジェクトについては実際には中断していないことが分かった。“Moved Repository” カテゴリに該当するプロジェクトは、何かしらの理由でリポジトリを移動し、移動先で開発が続いているプロジェクトである。例えば、refinerycms プロジェクトは“resolve/refinerycms”から“refinery/refinerycms”にリポジトリを移動している。“Keep Working” カテゴリに該当するプロジェクトは、調査中にGHTorrent データが作成された日である2015年4月1日以降にコミットが存在することを発見プロジェクトである。

中断していると判断したプロジェクトのうち、5プロジェクトはプロジェクトリーダーが辞めたことによる中断であった。このカテゴリでは、表2のasi-http-request プロジェクトの“I am no longer working on this library.”の様に、プロジェクトリーダーが辞める意思を表明していた。

また、他の理由として、依存している技術が使えなくなったというプロジェクトや、オープンソースからクローズドソースに移行したプロジェクト、開発計画がキャンセルされたプロジェクトなどがあつた。

そして、44プロジェクトについては、明示的な理由が見つからなかったため“Unknown”カテゴリとなっている。感情分析：今回、感情分析の対象とした10プロジェクトの情報を表3に示す。先に述べたとおり、中断した5プロ

ジェクトはプロジェクトリーダーが離脱したプロジェクトであり、継続している5プロジェクトは全ての継続しているプロジェクトからランダムに選択した。LOCの計測にはcloc^{*11}を、プログラミング言語の判定にはLinguist^{*12}を用いた。また、開発者数はコミット履歴から算出し、Starの数はGitHub ページを確認した^{*13}。

図2に中断したプロジェクトに対する感情分析の結果を示す。各図において、x軸は時系列を表し、y軸はコミットコメントの感情分析の値を表している。赤色の線はコミット毎の値の推移を表し、青色の線はLocal Regression (LOESS) を用いて平滑化した曲線であり、グレーの範囲は95%信頼区間である。同様に、図3に継続しているプロジェクトに対する感情分析の結果を示す。

図から分かる通り、全てのプロジェクトにおいて平滑化した曲線は0付近を通っており、コミットコメントは非常にニュートラルであることが分かる。加えて、中断したプロジェクトにおいて、後半ネガティブな傾向を表すことも無かつた。また、各コミットの値に着目しても、一部のコミットで3や-3の値を取るものの多くのコミットの値は[-2, 2]であり、それほど強い感情を示していない。

また、プロジェクトの前半と後半において、スコアに差があるかを確かめるため、コミットを日付で前半・後半の2つの期間に分けマン・ホイットニーのU検定を用いて2つのグループのコミットコメントの値に差があるかを確かめた。特に、本稿では後半に値が下がると仮定しているため、前半の値が後半の値よりも大きいということを片側検定を用いて確かめる。表4に結果を示す。表から分かる通り、全ての中断したプロジェクトにおいてニュートラルなコミットが最も多くなっている。また、マン・ホイットニーのU検定の結果では、“machinist”プロジェクトでp値が0.05を下回っているものの、他の4プロジェクトでは後半の値の方がネガティブであるとは言えなかつた。中断したプロジェクトと同様に、継続しているプロジェクトにおいても、ニュートラルなコミットの数が増えている。

中断したプロジェクト、継続しているプロジェクトに関わらず、コミットコメントはニュートラルな傾向を示す。また、プロジェクトの後半のコミットコメントの値の方がネガティブとは言えない。

5. 議論

本章では、本稿で得られた結果について議論する。

5.1 コミットコメントによる予測

本稿では、感情分析の対象としてコミットコメントを用

^{*10} プロジェクト期間が2012/01/01 - 2014/12/31であった場合、2013/09/01を境にプロジェクトを分ける。そのため、コミット数は必ずしも前半と後半で一致しない。

^{*11} <http://cloc.sourceforge.net/>

^{*12} <https://github.com/github/linguist>

^{*13} 2016年4月1日にアクセス

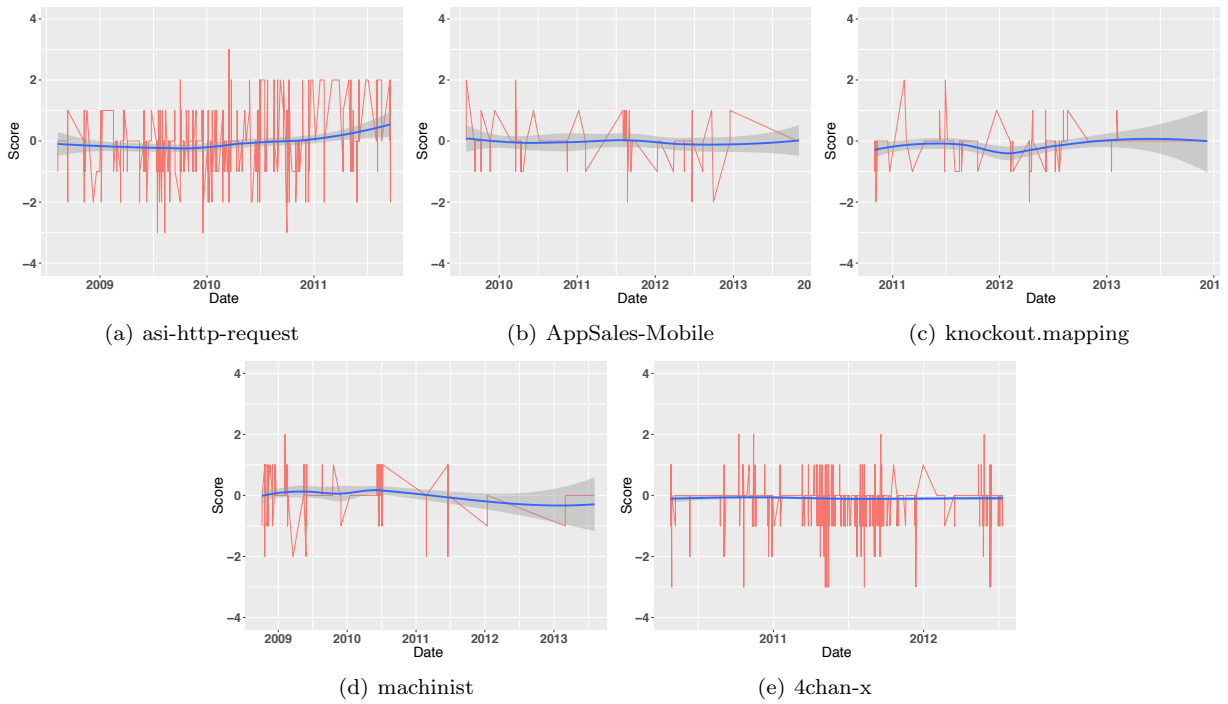


図 2 プロジェクトリーダーのコミットコメントの感情の推移 (中断プロジェクト)

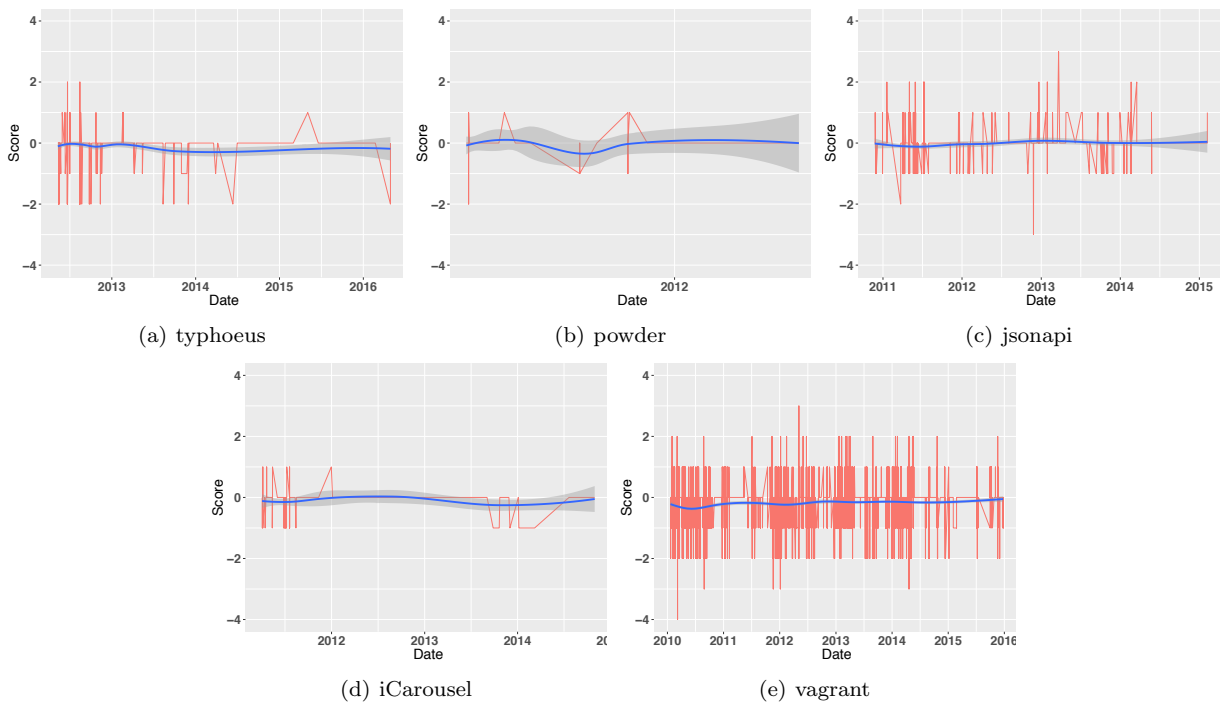


図 3 プロジェクトリーダーのコミットコメントの感情の推移 (継続プロジェクト)

いた。その結果、コミットコメントにおける感情はプロジェクト開始当初から中断前に至るまで、おおよそニュートラルな状態であることが分かった。コミットコメントが全体的にニュートラルな傾向を持つことは、Guzman ら [7] も指摘しているが、本稿ではプロジェクトリーダーが中断直前になってもニュートラルな傾向を持つことを示した。コミットコメントがニュートラルな傾向を持つことから、コミットコメントはその変更がどのようなものであるかを簡

潔に示す様にしており、そのため、コミットコメントには感情的な文章は出にくいのではないかと考えられる。

この結果から、本研究の最終目的である、中断するプロジェクトの予測にはコミットコメントは利用できないと考えられる。今後は、より自由な記述がされているであろう、pull request などへのレビューコメントや issue に関する議論に対して同様に感情分析を行い、プロジェクト中断の予測に利用できるか調査を行いたい。

表 4 感情分析結果

カテゴリ	プロジェクト名	感情分析 (コミット数)			U 検定
		ポジティブ	ニュートラル	ネガティブ	p 値
Terminated	asi-http-request	93	345	150	0.9745
	AppSales-Mobile	17	82	21	0.2712
	knockout.mapping	9	100	31	0.9618
	machinist	45	136	26	0.01227
	4chan-x	45	1,421	169	0.129
Survived	typhoeus	17	414	40	0.6176
	powder	4	34	4	0.5142
	jsonapi	56	474	95	0.9168
	iCarousel	6	99	23	0.1764
	vagrant	268	3,771	1,168	1.0

5.2 中断したプロジェクト

本稿では、コミット状況から中断していると考えられる 94 プロジェクトに対して、著者のうち一人が目視で各種情報 (GitHub ページ, オフィシャルページ, メーリングリストなど) を確認をすることにより、中断した原因を決定していった。リポジトリの移動などによって、継続していると判断したプロジェクトを含め、50 プロジェクトに関しては情報を確認することにより原因が判明したが、残りの 44 プロジェクトに関しては原因について記述を見つけることができなかつた。つまり、およそ半数のプロジェクトが明示的なアナウンスなく開発を中断しているということである (明示的なアナウンスがなくとも、pull request に対応をしない、メンションに反応しないなど、プロジェクトリーダーが活動を行っていないことが原因と思われるプロジェクトは見受けられたが、今回は Unknown としている)。

この状況を実際の OSS を利用した開発に即して考えると、明示的なアナウンスがある場合は利用しているプロジェクトが中断するにせよ、次の候補を考えるなどのアクションを取ることができる。しかし、明示的なアナウンスがない場合は、中断していることに気づくことなくその OSS を利用し続け、問題が発生した際に保守を受けることができず困るという可能性がある。そのため、このような明示的なアナウンスなく開発が中断するようなプロジェクトの中断を予測できるようになることは非常に有用であると考えられる。

6. 妥当性への脅威

6.1 構造的妥当性

本稿での、中断プロジェクトの決定はプロジェクトの活動によって行われている。SourceForge のようにプロジェクトの状態が開発者自身によって提供されている場合はその情報を用いることが最も適切だが [10], GitHub ではそのようなプロジェクトの状態に関する情報は提供されていない。そのため、本稿ではプロジェクトの活動を用いて、中断プロジェクトの決定を行った。プロジェクトの活動を

用いた方法は多くの論文 [3][4][19][24] で用いられており、この方法は適切だと考える。

6.2 内的妥当性

最も重要な内的妥当性への脅威は、中断理由の判断である。本稿では、著者の一人が理由について調査をし、判断を行った。この手法は、著者の主観のみに依存するものである。そのため、我々はデータ *14 を公開し、それを用いた検証実験が行われ、我々の結果を立証することを望む。

6.3 外的妥当性

本稿の結果は、プロジェクトリーダーが辞めたことによって中断した 5 プロジェクトに依存しており、一般化できるものではない。また、今回対象としたプロジェクトのプログラム言語は 4 種類 (Objective-C, JavaScript, Ruby, Java) と少なく、プログラム言語はプロジェクトに影響を与えることから [18], より大きなデータセットでの実験が結果の立証には必要である。

7. おわりに

本稿では、既存研究では難しいと考えられる、開発者が興味を失いプロジェクトを離脱するような状況を予測するために、感情分析を用いて予測することを試みた。離脱した場合、最もプロジェクトに対する影響が大きいことから、特にプロジェクトリーダーに着目し、プロジェクトリーダーによるコミットのコメントに対し感情分析を行い、プロジェクト開始当初から中断直前までの感情の推移を観測した。

その結果、中断直前の時期においても、既存研究 [7] が指摘した通り、コミットコメントはニュートラルな傾向を示した。また、時系列で中断したプロジェクトを前半・後半に分け、感情を比較したが差は無かつた。

今後は、コミットコメントのみならず、pull request や issue に対するコメントなどに対して分析対象を拡張し、プ

*14 <http://pos1.ait.kyushu-u.ac.jp/Disclosure/ses2016-sentimentanalysis.html>

プロジェクト中断の予測に利用できるかを検証したい。また、本稿のようにプロジェクトリーダーのみに着目するのではなく、core developer [13][14][27] のような、プロジェクト内で重要な役割を果たしている開発者にも着目をする必要があると考える。

参考文献

- [1] Bollen, J., Mao, H. and Zeng, X.: Twitter mood predicts the stock market, *CoRR*, Vol. abs/1010.3003 (2010).
- [2] Bonaccorsi, A. and Rossi, C.: Comparing motivations of individual programmers and firms to take part in the open source movement: From community to business, *Knowledge, Technology & Policy*, Vol. 18, No. 4, pp. 40–64 (2006).
- [3] Chengalur-Smith, S. and Sidorova, A.: Survival of Open-Source Projects: A Population Ecology Perspective, *Proc. of Int'l Conf. on Information Systems* (2003).
- [4] Chengalur-Smith, S., Sidorova, A. and Daniel, S. L.: Sustainability of Free/Libre Open Source Projects: A Longitudinal Study., *Journal of the Association for Information Systems*, Vol. 11, No. 11 (2010).
- [5] Evangelopoulos, N., Sidorova, A., Fotopoulos, S. and Chengalur-Smith, I.: Determining Process Death Based on Censored Activity Data, *Communications in Statistics - Simulation and Computation*, Vol. 37, No. 8, pp. 1647–1662 (2008).
- [6] Gousios, G.: The GHTorrent dataset and tool suite, *Proc. Int'l Working Conf. on Mining Software Repositories (MSR)*, pp. 233–236 (2013).
- [7] Guzman, E., Azócar, D. and Li, Y.: Sentiment Analysis of Commit Comments in GitHub: An Empirical Study, *Proc. Working Conf. on Mining Software Repositories (MSR)*, pp. 352–355 (2014).
- [8] Guzman, E. and Bruegge, B.: Towards Emotional Awareness in Software Development Teams, *Proc. Joint Meeting on Foundations of Software Engineering (ESEC/FSE)*, pp. 671–674 (2013).
- [9] Kalliamvakou, E., Gousios, G., Blincoe, K., Singer, L., German, D. M. and Damian, D.: The Promises and Perils of Mining GitHub, *Proc. Working Conf. on Mining Software Repositories (MSR)*, pp. 92–101 (2014).
- [10] Khondhu, J., Capiluppi, A. and Stol, K.-J.: Is It All Lost? A Study of Inactive Open Source Projects, *Proc. Int'l Conf. on Open Source Systems*, pp. 61–79 (2013).
- [11] Michlmayr, M.: Community Management in Open Source Projects, *The European Journal for the Informatics Professional*, Vol. X, No. 3, pp. 22–26 (2009).
- [12] Mishra, N. and Jha, C. K.: Article: Classification of Opinion Mining Techniques, *International Journal of Computer Applications*, Vol. 56, No. 13, pp. 1–6 (2012).
- [13] Mockus, A., Fielding, R. T. and Herbsleb, J. D.: Two Case Studies of Open Source Software Development: Apache and Mozilla, *ACM Trans. on Software Engineering and Methodology*, Vol. 11, No. 3, pp. 309–346 (2002).
- [14] Nakakoji, K., Yamamoto, Y., Nishinaka, Y., Kishida, K. and Ye, Y.: Evolution Patterns of Open-source Software Systems and Communities, *Proc. Int'l Workshop on Principles of Software Evolution (IWPSE)*, pp. 76–85 (2002).
- [15] Ortu, M., Adams, B., Destefanis, G., Tourani, P., Marchesi, M. and Tonelli, R.: Are Bullies More Productive?: Empirical Study of Affectiveness vs. Issue Fixing Time, *Proc. Working Conf. on Mining Software Repositories (MSR)*, pp. 303–313 (2015).
- [16] Pletea, D., Vasilescu, B. and Serebrenik, A.: Security and Emotion: Sentiment Analysis of Security Discussions on GitHub, *Proc. Working Conf. on Mining Software Repositories (MSR)*, pp. 348–351 (2014).
- [17] Raja, U. and Tretter, M.: Defining and Evaluating a Measure of Open Source Project Survivability, *IEEE Transactions on Software Engineering*, Vol. 38, No. 1, pp. 163–174 (2012).
- [18] Ray, B., Posnett, D., Filkov, V. and Devanbu, P.: A Large Scale Study of Programming Languages and Code Quality in Github, *Proc. Int'l Symposium on Foundations of Software Engineering (FSE)*, pp. 155–165 (2014).
- [19] Samoladas, I., Angelis, L. and Stamelos, I.: Survival Analysis on the Duration of Open Source Projects, *Inf. Softw. Technol.*, Vol. 52, No. 9, pp. 902–922 (2010).
- [20] Thelwall, M.: Heart and Soul: Sentiment Strength Detection in the Social Web with SentiStrength 1.
- [21] Thelwall, M., Buckley, K. and Paltoglou, G.: Sentiment strength detection for the social web, *Journal of the American Society for Information Science and Technology*, Vol. 63, No. 1, pp. 163–173 (2012).
- [22] Tourani, P. and Adams, B.: The Impact of Human Discussions on Just-In-Time Quality Assurance, *Proc. Int'l Conf. on Software Analysis, Evolution, and Reengineering (SANER)* (2016).
- [23] Tourani, P., Jiang, Y. and Adams, B.: Monitoring Sentiment in Open Source Mailing Lists: Exploratory Study on the Apache Ecosystem, *Proc. Annual Int'l Conf. on Computer Science and Software Engineering (CASCON)*, pp. 34–44 (2014).
- [24] Wang, J.: Survival factors for Free Open Source Software projects: A multi-stage perspective, *European Management Journal*, Vol. 30, No. 4, pp. 352 – 371 (2012).
- [25] Wilson, T., Wiebe, J. and Hoffmann, P.: Recognizing Contextual Polarity in Phrase-level Sentiment Analysis, *Proc. Conf. on Human Language Technology and Empirical Methods in Natural Language Processing (HLT)*, pp. 347–354 (2005).
- [26] Yamashita, K., Kamei, Y., McIntosh, S., Hassan, A. E. and Ubayashi, N.: Magnet or Sticky? Measuring Project Characteristics from the Perspective of Developer Attraction and Retention, *Journal of Information Processing*, Vol. 24, No. 2, pp. 339–348 (online), DOI: 10.2197/ipsjip.24.339 (2016).
- [27] Yamashita, K., McIntosh, S., Kamei, Y., Hassan, A. E. and Ubayashi, N.: Revisiting the Applicability of the Pareto Principle to Core Development Teams in Open Source Software Projects, *Proc. Int'l Workshop on Principles of Software Evolution (IWPSE)*, pp. 46–55 (2015).