

論文紹介

Sentiment Analysis for Software Engineering: How Far Can We Go?

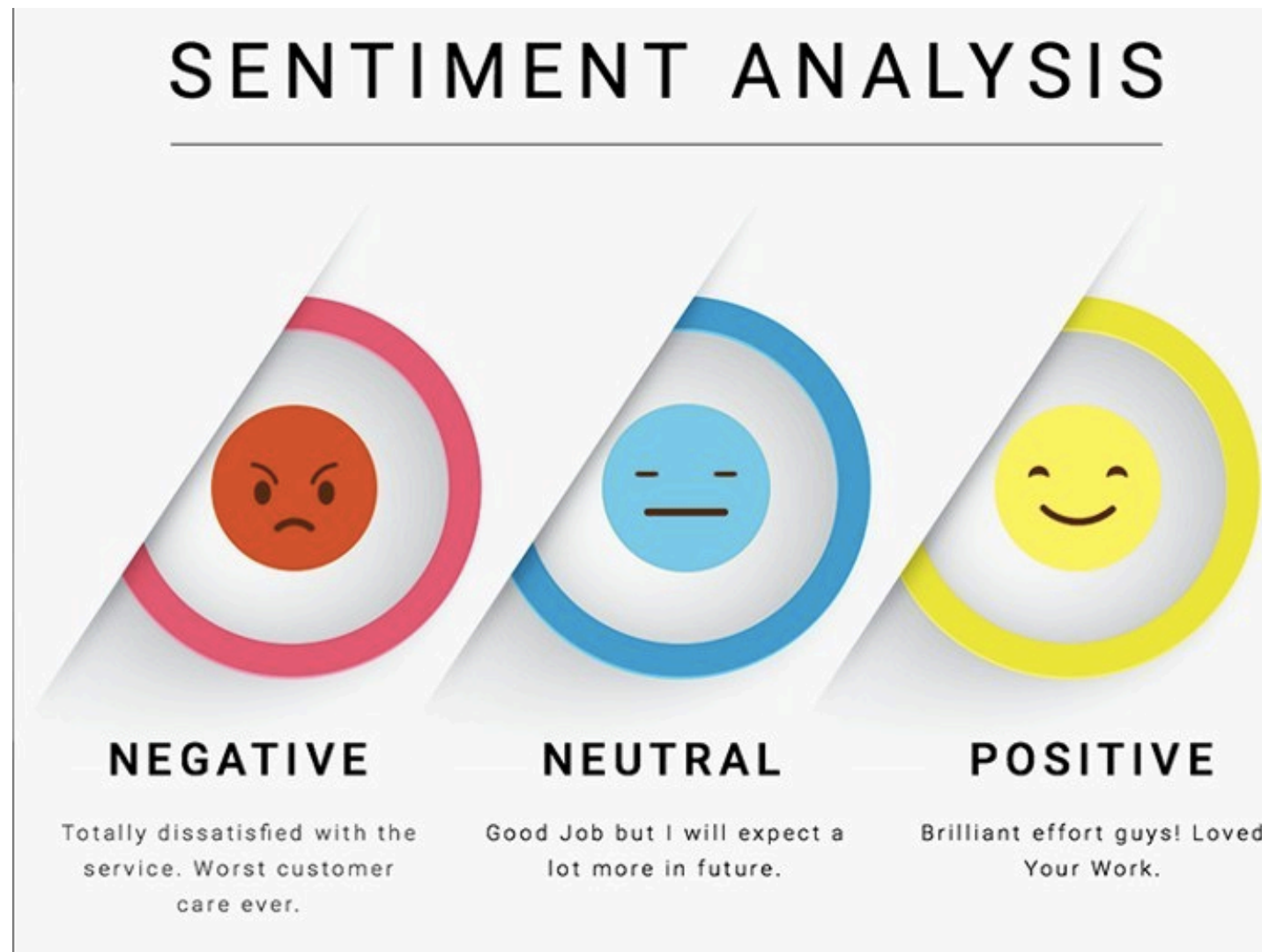
Bin Lin
Fiorella Zampetti
Gabriele Bavota
Massimiliano Di Penta
Michele Lanza
Rocco Oliveto

背景

- 近年、意見などのテキストデータの自動分析に感情分析ツールが利用されている
 - 例：映画の視聴者の意見の要約
- 感情分析ツールはSE（Software Engineering）分野でも取り入れられている
 - AppReviewの分析
 - 開発チームのdistress, happinessの測定
 -

感情分析

- Sentiment Analysis
 - 書き手の感情を分析する手法



感情分析に関する研究

- 一般的に利用される sentimental tool は SentiStrength
- SE テキスト 向けに改良された SentiStrength-SE もある

Table 1: Sentiment analysis tools used for SE applications.

Tool	Technique	Trained on	Used by
SentiStrength	Rule-based	MySpace	[7–11] [15, 22, 24] [31, 33] [36]
NLTK/VADER	Rule-based	Micro-Blog	[30] [28]
Stanford CoreNLP	Recurs. Neural Net	Movie Reviews	[29], our work
EmoText	Lexical Features	Stack Overflow, JIRA	[2]
SentiStrength – SE	SentiStrength	JIRA	[15]
Uddin and Khomh	Sentim. Orientation [13]	API Reviews	[36]

問題と目的

- 問題

- 感情分析ツールは，ソフトウェア関連のテキストを扱うように設計されいないものが多い
- 学習データと別ドメインに適用すると精度が落ちると言われている

- 目的

- 感情分析ツールをSEテキストに適用した時に，どの程度精度に影響を与えるか？
- 学習データによる影響はあるのか？

アプローチ

1. 学習データによる精度の変化を確認するため

- StackOverflowのデータを学習させたStanford CoreNLP SOを作成

2. 感情分析ツールの精度比較のため

- Stack Overflow, App reviews, JIRA issuesに各感情分析ツールを適用

・ 手動によりラベルづけされたデータセット

Table 4: Dataset used for evaluating sentiment analysis tools in software engineering

Dataset	# sentences	# positive	# neutral	# negative
Stack Overflow	1,500	178	1,191	131
App reviews	341	186	25	130
JIRA issue	926	290	0	636

結果

dataset	tool	# correct prediction	positive precision	positive recall	neutral precision	neutral recall	negative precision	negative recall
Stack Overflow	SentiStrength	1,043	0.200	0.359	0.858	0.772	0.397	0.433
	NLTK	1,168	0.317	0.244	0.815	0.941	0.625	0.084
	Stanford CoreNLP	604	0.231	0.344	0.884	0.344	0.177	0.837
	SentiStrength-SE	1,170	0.312	0.221	0.826	0.930	0.500	0.185
	Stanford CoreNLP SO	1,139	0.317	0.145	0.836	0.886	0.365	0.365
App reviews	SentiStrength	213	0.745	0.866	0.113	0.320	0.815	0.338
	NLTK	184	0.751	0.812	0.093	0.440	1.000	0.169
	Stanford CoreNLP	237	0.831	0.715	0.176	0.240	0.667	0.754
	SentiStrength-SE	201	0.741	0.817	0.106	0.400	0.929	0.300
	Stanford CoreNLP SO	142	0.770	0.253	0.084	0.320	0.470	0.669
JIRA issues	SentiStrength	714	0.850	0.921	-	-	0.993	0.703
	NLTK	276	0.840	0.362	-	-	1.000	0.269
	Stanford CoreNLP	626	0.726	0.621	-	-	0.945	0.701
	SentiStrength-SE	704	0.948	0.883	-	-	0.996	0.704
	Stanford CoreNLP SO	333	0.635	0.252	-	-	0.724	0.409

- Stanford CoreNLP SOでは、ほぼ同じ精度を維持しつつ良いRecallを示している（と書かれてるけどそうは見えない…）
- Stanford CoreNLP SOの精度が低い理由として、SOの学習データが少なかったことが考えられる

結果

dataset	tool	# correct prediction	positive precision	positive recall	neutral precision	neutral recall	negative precision	negative recall
Stack Overflow	SentiStrength	1,043	0.200	0.359	0.858	0.772	0.397	0.433
	NLTK	1,168	0.317	0.244	0.815	0.941	0.625	0.084
	Stanford CoreNLP	604	0.231	0.344	0.884	0.344	0.177	0.837
	SentiStrength-SE	1,170	0.312	0.221	0.826	0.930	0.500	0.185
	Stanford CoreNLP SO	1,139	0.317	0.145	0.836	0.886	0.365	0.365
App reviews	SentiStrength	213	0.745	0.866	0.113	0.320	0.815	0.338
	NLTK	184	0.751	0.812	0.093	0.440	1.000	0.169
	Stanford CoreNLP	237	0.831	0.715	0.176	0.240	0.667	0.754
	SentiStrength-SE	201	0.741	0.817	0.106	0.400	0.929	0.300
	Stanford CoreNLP SO	142	0.770	0.253	0.084	0.320	0.470	0.669
JIRA issues	SentiStrength	714	0.850	0.921	-	-	0.993	0.703
	NLTK	276	0.840	0.362	-	-	1.000	0.269
	Stanford CoreNLP	626	0.726	0.621	-	-	0.945	0.701
	SentiStrength-SE	704	0.948	0.883	-	-	0.996	0.704
	Stanford CoreNLP SO	333	0.635	0.252	-	-	0.724	0.409

- Positive/NegativeとNeutralの区別が難しい
 - StackOverflowのように感情が見えにくいデータではNeutralの精度が高い（positive/negativeは識別が難しい）
 - AppReviewのような感情が見えやすいデータではpositive/negativeの精度が高い

結果

dataset	tool	# correct prediction	positive precision	positive recall	neutral precision	neutral recall	negative precision	negative recall
Stack Overflow	SentiStrength	1,043	0.200	0.359	0.858	0.772	0.397	0.433
	NLTK	1,168	0.317	0.244	0.815	0.941	0.625	0.084
	Stanford CoreNLP	604	0.231	0.344	0.884	0.344	0.177	0.837
	SentiStrength-SE	1,170	0.312	0.221	0.826	0.930	0.500	0.185
	Stanford CoreNLP SO	1,139	0.317	0.145	0.836	0.886	0.365	0.365
App reviews	SentiStrength	213	0.745	0.866	0.113	0.320	0.815	0.338
	NLTK	184	0.751	0.812	0.093	0.440	1.000	0.169
	Stanford CoreNLP	237	0.831	0.715	0.176	0.240	0.667	0.754
	SentiStrength-SE	201	0.741	0.817	0.106	0.400	0.929	0.300
	Stanford CoreNLP SO	142	0.770	0.253	0.084	0.320	0.470	0.669
JIRA issues	SentiStrength	714	0.850	0.921	-	-	0.993	0.703
	NLTK	276	0.840	0.362	-	-	1.000	0.269
	Stanford CoreNLP	626	0.726	0.621	-	-	0.945	0.701
	SentiStrength-SE	704	0.948	0.883	-	-	0.996	0.704
	Stanford CoreNLP SO	333	0.635	0.252	-	-	0.724	0.409

- JIRAに関しては、Positive/Negativeの分類だけのため精度が高い

まとめ

- 感情分析ツールをSEテキストに大使適用した際の精度の影響を調査した
 - Positive/NegativeとNeutralの判別が難しい
- 精度に関しては、テキストによっては問題がある可能性もある

所感

- 良かった点
 - 感情分析ツールを比較した点が面白い
 - SEテキスト向けのツールがあることを初めて知った
- 疑問点
 - CoreNLP SOの学習データセットをCoreNLPと同じだけにすると精度はどうなるのか？
 - Positive/Negative/Neutralのテキストの数がかなり違うけどいいのか？