

Exploring Word Embedding Techniques to Improve Sentiment Analysis of Software Engineering Texts

Eeshita Biswas, K. Vijay-Shanker, Lori Pollock

発表者：作島大智

Sentimental Analysis

- 表現されている意見が肯定的, 否定的, または中立的であるかどうかを判断するためにテキストの単位を分析する自然言語処理のマイニング手法
- 以下, SAと表記

目的

SEアーティファクトの感情分析のためのトレーニングを改善する2つの潜在的な要素の影響を調査すること

- データの不均衡
- SAツールのカスタマイズ

背景

- 感情分析は、SOなどのQ&Aサイトにおけるソースコード改善のための洞察に満ちたコメントを推奨するなどの研究が行われており、SEドメインに適応された既存の感情分析ツールの分析が行われている。
- 肯定的、否定的、中立的な意見に関するデータセットが**不均衡**である
- トレーニングデータにニュースや映画のレビューなどノイズがある
- 先行研究 [25]でSEドメイン用にドメイン固有のSAツールに**カスタマイズ**することで精度が向上する

Linら[19]の研究(1/2)

- 既存のツールをソフトウェア分野への適応を検討
→適切なトレーニングデータを使用してSAツールをドメイン固有なものにカスタマイズする
- 既存の感情分析ツールの分析にはSVMなどの特徴機械学習手法が主であるが、Linらは再帰的ニューラルネットワーク法を用いた

Linら[19]の研究(2/2)

- SOからJavaタグでlibrary/libraries, API(s)のいずれかの単語を含む全ての会話リストを抽出
- Stanford CoreNLPツールキットを使って507万のセンテンスを抽出
- この文章から1500のセンテンスを無作為に選択
- 1500のセンテンスを手作業でラベル付け
- 否定的178文, 中立的1191文, 肯定的131文

12%

80%

8%
- 10倍交差検定を行い, 1500の文全てを使用

Linら[19]の研究の問題点

- 全体的な精度は高いが、否定的、肯定的な文のprecisionとrecallが低い
- **中立的な文を分類できているだけ**
- 一定のニューラル分類器でも同じだけの精度を出すことが可能

Research Question

- 1) データセット分布の不均衡は、SEアーティファクトへの感情分析の適用において満足な結果が得られない要因にどの程度なっているのか.
- 2) ドメイン固有の単語埋め込みを使用してソフトウェアドメインに対する感情分類子をさらにカスタマイズすると、感情分析プロセスにどの程度の影響があるのか.

RQ1 : アプローチ

データの取得(1/2)

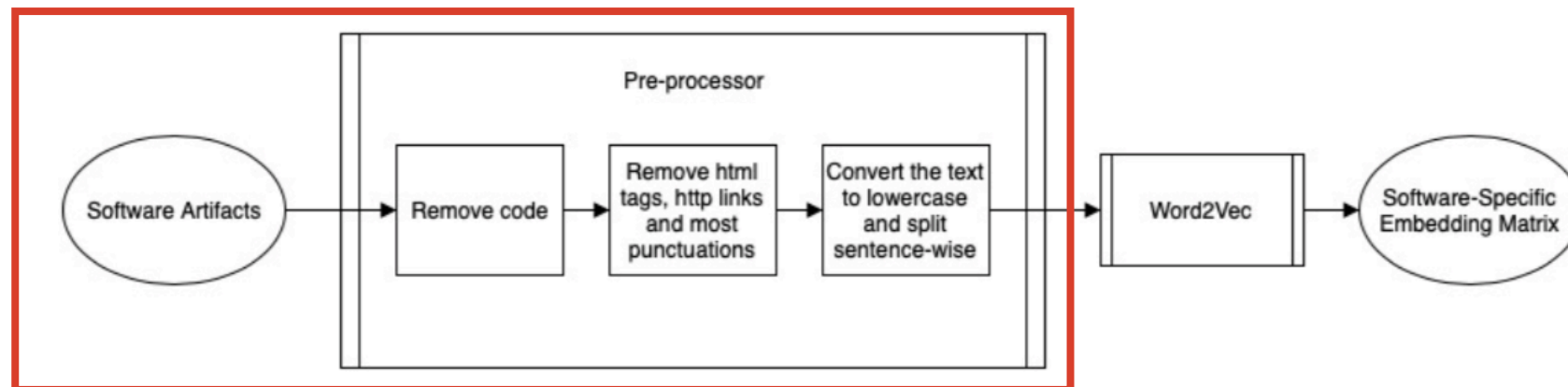


Fig. 1. Creating a software-specific word embedding matrix as by Efstathiou et al.[8]

- SOのソフトウェア関連のテキスト(“0808~”1806)を使用
- 質問と回答のみ使用してデータセットを構築
- 前準備として、コードセグメント削除、HTMLタグとHTTPリンクの削除、単語を全て小文字で統一する
- ストップワードの削除は行わない
→文の意味合いが変化するため

データの取得(2/2)

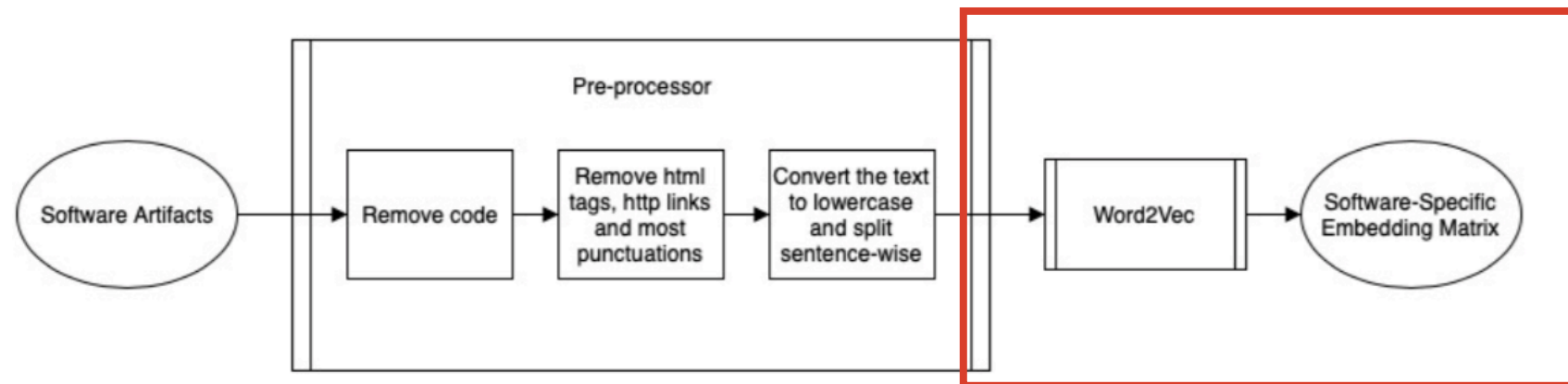


Fig. 1. Creating a software-specific word embedding matrix as by Efstathiou et al.[8]

- Word2VecのPythonライブラリGensimを使用して, word embedding を作成
- 前処理から文をトークン化し, トークン化されたリストをWord2Vecの入力とする
- Efstathiou [8]で使用されていると同じパラメータを使用
- 200個の次元と180万の単語の語彙

サンプリング手法

- ・ **オーバーサンプリング**

複製もしくは、合成例(synthetic examples)を作ること
で少数クラスを**増やす**手法(SMOTE [5]など)

- ・ **アンダーサンプリング**

クラス分布を変更するために多数派クラスの例を**破棄する**

サンプリング手法

- **オーバーサンプリング**は少数クラスのデータの複製や合成
→多数派クラス, 少数派クラス共にデータ例を全て保持することが可能であるが, 訓練データの重複を行うため, **過学習**を行なってしまう.
- **アンダーサンプリング**は多数派クラスのデータの破棄
→クラスの不均衡は解消し精度は向上するが
データセットが大幅に減少する

データセット構築

- Linらが用いたSOから取得した1500の文を使用
- 1500の文を10のグループに分ける
- 10のグループで3つのクラスがすべて均等な分布を維持
- 1グループをテストセット, 9グループは訓練セット
- 訓練セットにのみサンプリングのための変更が行う

サンプリング率

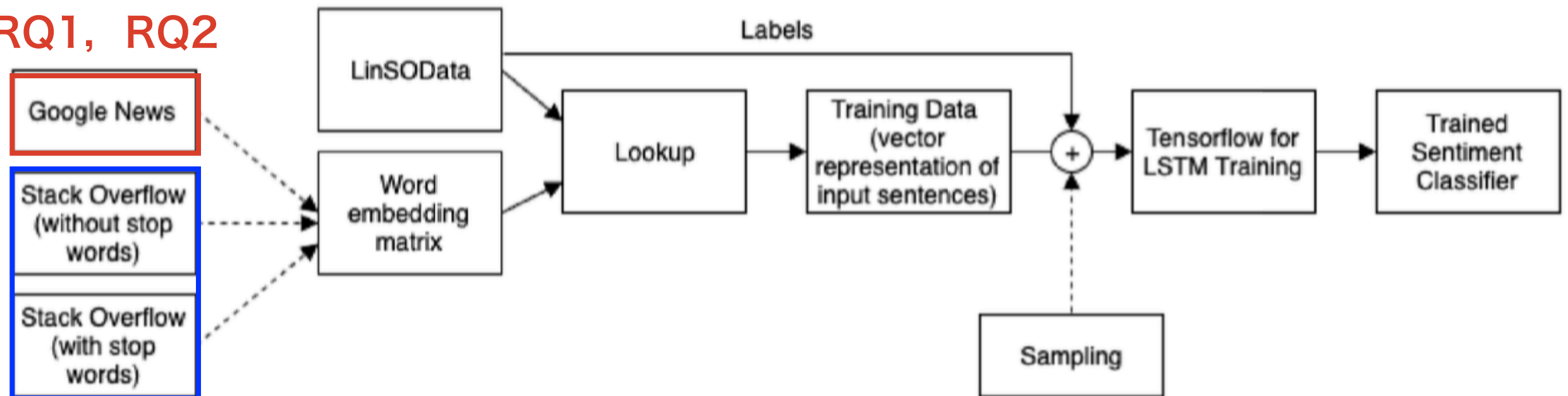
- (neutral , negative , positive)
- (1,7,9) : 全てのクラスが均等に分布, 元のデータ分布
- (1,5,6) : バランスはできているが均等分布ではない
- (0.8,6,7) :
全てのクラスが均等に分布, 訓練データの8割を保持
- (0.6,4,5) :
全てのクラスが均等に分布, 訓練データの6割を保持

感情分類子のための ニューラルネットワークアーキテクチャの構築

文章のベクトル表現の構築

トレーニング

RQ1, RQ2



RQ2

Fig. 2. Training the Sentiment Classifier.

ラベル付けにはワンホット表現を使用

[1,0,0] : negative

[0,1,0] : neutral

[0,0,1] : positive

著者らのモデル

- 一般的な単語埋め込みモデル
- Google News Word Embedding [34] を感情分類子の入力とした
- Linらと同様に10倍交差検定を行う

結果

RQ1

1) データセット分布の不均衡は, SEアーティファクトへの感情分析の適用において満足な結果が得られない要因にどの程度なっているのか.

RQ1：結果

著者ら

TABLE I
TESTING RESULTS OF OUR SENTIMENT CLASSIFIER WITH DIFFERENT SAMPLING TECHNIQUES FOR DATA IMBALANCE.

Sampling Rates	Correct Prediction	Negative		Neutral		Positive	
		Precision	Recall	Precision	Recall	Precision	Recall
Original	1200	0.453	0.433	0.881	0.908	0.402	0.313
1, 7, 9	1171	0.43	0.517	0.884	0.861	0.455	0.42
1, 5, 6	1137	0.47	0.512	0.888	0.832	0.32	0.458
0.8, 6, 7	1160	0.43	0.562	0.801	0.734	0.466	0.42
0.6, 4, 5	1134	0.4	0.53	0.88	0.821	0.37	0.413

Linら

TABLE II
TESTING RESULTS FROM ‘SENTIMENT ANALYSIS FOR SOFTWARE ENGINEERING: HOW FAR CAN WE GO?’ [19]

Correct Prediction	Negative		Neutral		Positive	
	Precision	Recall	Precision	Recall	Precision	Recall
1139	0.365	0.365	0.836	0.886	0.317	0.145

0.2UP

0.3UP

- データの不均衡を完全することで精度が向上した
- オーバーサンプリングおよびアンダーサンプリングは少数派クラスに対する感情分析子の精度を積極的に向上させた。

RQ1：結果

- 少数派クラスに対して非常に大きい割合でオーバーサンプリングしたため結果として過学習する傾向にあるのでは
- オーバーサンプリング率を下げるとデータの不均衡が起こる
→データの不均衡に着目しているため、仕方がない
- アンダーサンプリング率をあげて中立的な意見を多く破棄した場合、データセットのサイズが元の半分未満になり合理的ではない
- オーバーのみとハイブリッドでは結果が同じである
→アンダーサンプリングのために起こった**情報の損失**が**過学習の減少**によって相殺されたと考えられる。

正しく識別された文の例

TABLE III
EXAMPLES OF MINORITY CLASS SENTENCES CLASSIFIED NEUTRAL ON ORIGINAL DATASET BUT CLASSIFIED ACCURATELY ON THE RESAMPLED DATASET USING THE GOOGLE NEWS WORD EMBEDDINGS

Sentence	Class
<i>A workaround might be to use reflection, however this isn't a generally recommended approach for a number of reasons (brittleness, performance problems, breaking encapsulation, etc).</i>	Negative
<i>I am using standout library, I added a videoView in it, Everything is running correctly, Except if i click on video then mediaController Is not Showing and i am getting null pointer exception.</i>	Negative
<i>Enabled and pointed my client to a secure port on my container and the last example (the one that causes the exception) works successfully.</i>	Positive
<i>If you are looking for a way to write to streams in a more intuitive way, try CODE_FRAGMENT.</i>	Positive

RQ2

2) ドメイン固有の単語埋め込みを使用してソフトウェアドメインに対する感情分類子をさらにカスタマイズすると、感情分析プロセスにどの程度の影響があるのか.

RQ2：アプローチ

- 専門のソフトウェアドメイン特有の知識のために SOwoStop word embedding [8]を使用
- GN word embedding を SOwoStop word embedding に置き換え, 10倍交差検定を行う.
- ソフトウェアドメインにカスタマイズせずに感情分類器と比較するために, RQ1の結果を使用した

RQ2 : 結果

RQ1 : 一般的なword embedding (Google News)

TABLE I

TESTING RESULTS OF OUR SENTIMENT CLASSIFIER WITH DIFFERENT SAMPLING TECHNIQUES FOR DATA IMBALANCE.

Sampling Rates	Correct Prediction	Negative		Neutral		Positive	
		Precision	Recall	Precision	Recall	Precision	Recall
Original	1200	0.453	0.433	0.881	0.908	0.402	0.313
1, 7, 9	1171	0.43	0.517	0.884	0.861	0.455	0.42
1, 5, 6	1137	0.47	0.512	0.888	0.832	0.32	0.458
0.8, 6, 7	1160	0.43	0.562	0.801	0.734	0.466	0.42
0.6, 4, 5	1134	0.4	0.53	0.88	0.821	0.37	0.413

ソフトウェア特有のword embedding(SOwoStop, ストップワードあり)

TABLE IV

TESTING RESULTS OF OUR SENTIMENT CLASSIFIER USING SOwoSTOP WORD EMBEDDINGS (WITHOUT STOP WORDS)

Sampling Rates	Correct Prediction	Negative		Neutral		Positive	
		Precision	Recall	Precision	Recall	Precision	Recall
Original	1172	0.35	0.216	0.839	0.913	0.40	0.32
1, 7, 9	1126	0.333	0.37	0.858	0.854	0.352	0.32
1, 5, 6	1177	0.39	0.38	0.865	0.875	0.47	0.39
0.8, 6, 7	1042	0.23	0.41	0.87	0.831	0.41	0.35
0.6, 4, 5	1101	0.32	0.40	0.88	0.84	0.345	0.48

RQ2：結果

RQ1：一般的なword embedding (Google News)

TABLE I

TESTING RESULTS OF OUR SENTIMENT CLASSIFIER WITH DIFFERENT SAMPLING TECHNIQUES FOR DATA IMBALANCE.

Sampling Rates	Correct Prediction	Negative		Neutral		Positive	
		Precision	Recall	Precision	Recall	Precision	Recall
Original	1200	0.453	0.433	0.881	0.908	0.402	0.313
1, 7, 9	1171	0.43	0.517	0.884	0.861	0.455	0.42
1, 5, 6	1137	0.47	0.512	0.888	0.832	0.32	0.458
0.8, 6, 7	1160	0.43	0.562	0.801	0.734	0.466	0.42
0.6, 4, 5	1134	0.4	0.53	0.88	0.821	0.37	0.413

ソフトウェア特有のword embedding(SOwoStop, ストップワードあり)

TABLE IV

TESTING RESULTS OF OUR SENTIMENT CLASSIFIER USING SOWoSTOP WORD EMBEDDINGS (WITHOUT STOP WORDS)

Sampling Rates	Correct Prediction	Negative		Neutral		Positive	
		Precision	Recall	Precision	Recall	Precision	Recall
Original	1172	0.35	0.216	0.839	0.913	0.40	0.32
1, 7, 9	1126	0.333	0.37	0.858	0.854	0.352	0.32
1, 5, 6	1177	0.39	0.38	0.865	0.875	0.47	0.39
0.8, 6, 7	1042	0.23	0.41	0.87	0.831	0.41	0.35
0.6, 4, 5	1101	0.32	0.40	0.88	0.84	0.345	0.48

ソフトウェア特有のword embedding(SOwoStop, ストップワードなし)

TABLE V

TESTING RESULTS OF OUR SENTIMENT CLASSIFIER USING SOWSTOP WORD EMBEDDINGS (WITH STOPWORDS)

Sampling Rates	Correct Prediction	Negative		Neutral		Positive	
		Precision	Recall	Precision	Recall	Precision	Recall
Original	1200	0.522	0.068	0.796	0.993	1	0.031
1, 7, 9	1205	0.5	0.6	0.893	0.881	0.442	0.38
1, 5, 6	1168	0.472	0.34	0.844	0.89	0.32	0.33
0.8, 6, 7	944	0.225	0.624	0.915	0.706	0.318	0.301
0.6, 4, 5	1038	0.41	0.53	0.92	0.746	0.226	0.54

RQ2：結果

- ソフトウェア固有のword embeddingによる感情分類子の追加のカスタマイズは一般的なword embeddingを使用した感情分類子によって達成された結果を上回るものではない
 - 分野特化にカスタマイズしても影響なしという結果
- しかし影響がないとは言えない
 - データのサイズや量などの単語埋め込みのトレーニングの際、条件パラメータがある可能性がある。

RQ2：考察

- Google News単語埋め込みははるかに大きなコーパスで訓練され，訓練されたモデルが300万単語のベクトルを持つ
- 専門家のみが書いているのでデータの品質が良い
- 1500文のデータセットにはソフトウェア固有のすべての文ではなく，一般的な英文が含まれている
- ソフトウェア特有のword embedding(SOwoStop)のデータ数がGoogle News と同程度あれば結果が異なった可能性がある。

制約

- Linらの研究でのラベルが手動で行われていること
- 最適な結果を得ることができるハイパーパラメータの構成が最適でない可能性

まとめ

- サンプルング率を調整すると、多数派クラスの予測に影響を与ることなく、Recallの観点から少数派クラスに対するモデルの精度が向上する.
- オーバーサンプリングとハイブリッド(オーバーとアンダー)を用いると、少数派クラスの否定的、肯定的な文を多数識別することが可能.
- GN word embeddingを使用する場合と SOwStopword embeddingを使用する場合のパフォーマンスは同程度であった.

所感

- 少数派クラスにおける大きな割合のオーバーサンプリングがもたらす結果の精度の高さの理由が過学習ではないと言い切れるのはなにで言えるのか
- どの程度データ数を増やせばカスタマイズの効果があるのかというのが気になった
- やはりデータ数を確保することが大変である

関連研究

関連研究

- Touraniら
- 主流の既成SAツールであるSentiStrengthを使用
- データはthe Apacheソフトウェア財団の2つのプロジェクト(Tomcat, Ant)のユーザ, 開発者メーリスから感情情報(sentiment information)を抽出
- Positive : 29.56%
Negative : 13.1%

関連研究

- Novielliら [23]
- 感情状態を分析するためにSA技法を採択することの議題 (性格特性からSOでの表現など)についての議論を行う.

関連研究

- Jongelingら [16]
- 広く使われている4つのSAツールの比較を行う.
SentiStrength, NLTK, Stanford CoreNLP,
AlchemyAPI
- どのツールもSEドメインで表現された感情を性格に予測
することはできていないと結論

関連研究

- Blaz, Becker [3]
- IT分野で感情を含む用語をまとめたドメイン辞書を作成
- ITサポートに含まれる感情を評価

関連研究

- Ahmedら
- OSSからのコードレビューコメント用にカスタマイズされたSAツール, SentiCR [1] を開発
- NLTK, TF-IDF機能, 教師あり学習アルゴリズムを使用

関連研究

- IslamとZibran [14]
- 構成パラメータを変更して heuristic な規則を追加した SentiStrengthの拡張版, SentiStrength-SE [14]を開発
- JIRA issue trackingからのissueコメントについて SentiStrength-SEを評価し, SentiStrengthを上回る改善を発見

関連研究

- IslamとZibran

価数と覚醒空間？

- JIRA issue trackingを使用して， valence and arousal spaceの両方で感情を自動分析するための辞書ベースの語彙アプローチであるDEVA [13]を発表

関連研究

- Ding
- GitHubプロジェクトから発行された一連の重要事項について，特徴ベクトル化と教師付き分類アルゴリズムを用いたエンティティレベルSAツールであるSentiSW [7]を構築

関連研究

- Calefato
- 感情分類子であるSenti4SD [4]を構築
- 手動で注釈をつけ, 3つのクラスすべてのバランスのとれた分布を持つSOの質問, 回答, およびコメントのgold standardを使用して訓練および検証を行う.
- 肯定的, 中立的意見を否定的意見とする誤分類を減らす

関連研究

- Novielli [25]
- SEデータについてSAツール：SentiStrength, SentiStrength-SE, Senti4SD, SentiCRの比較
- SE固有のカスタマイズは正確さを向上させると結論

関連研究

- Imtiaz [11]
- 既存の感情と丁寧さの検出ツールに関する調査を実施
- 既存のツールは信頼性が低いと判断し，開発者の討論における感情と丁寧さを識別できていない

関連研究

- Islam and Zibran
- ソフトウェアドメイン固有のSAツールについて比較研究を実施
- データセットごとに異なるツールの精度を発見
→データによって有効なツールが違う？

関連研究

- Xuら [36]
- Word embedding と SMOTE を使用して完全にバランスの取れたデータセットを作成するWEC-Moteという手法を提案
- study[18,20]
under, under&overの組み合わせ[20]が感情分析の分類性能をどのように改善するのかを示す.