複数の表現学習手法を用いた日本語ツイートの感情強度推定

Estimation of Emotion Intensities in Japanese Tweets using Representation Learning

佐藤 一輝[1]　尾崎 知伸[2]*
Kazuki Sato[1]　Tomonobu Ozaki[2]

[1] 日本大学 大学院 総合基礎科学研究科
[1] Graduate School of Integrated Basic Sciences, Nihon University
[2] 日本大学文理学部
[2] College of Humanities and Sciences, Nihon University

Abstract: Emotion estimation of tweets is expected to be applied in various fields such as opinion analysis and financial prediction. To improve the applicability of emotion estimation, in this paper, we consider the estimation tasks for emotion intensities. We first prepare a database of Japanese tweets annotated for pleasant, anger, sadness and joy intensities using best-worst scaling, and then build various models for estimating the intensities of emotion. Three algorithms on representation learning are employed to build the estimation models, and are examined experimentally. As a result, we confirm that a certain model can not only classify the emotions but also predict intensities accurately.

1　はじめに

Twitter は，SNS 上でツイートと呼ばれる 140 文字以内の短文を投稿・閲覧するサービスである．その手軽さから，利用者の感情や意見が現れやすい傾向があると考えられ，これまでに，選挙などのイベントに対する意見の分析や，製品の評判分析，株価の予測など，ツイートを対象とした感情種の推定に関する応用が数多く行われている．また近年では，感情の種類だけではなく，その強さ（感情強度）を推定する研究が行われている．例えば，WASSA-2017[1] や SemEval 2018: Task1 Affect in Tweets[2] では，Anger，Fear，Joy，Sadness の 4 感情を対象に，ツイートから各感情の強度（0.0 ～ 1.0）を推定するタスクが設定されている．

これらの研究の多くは英文ツイートを対象としたものであり，日本語ツイートに関しては必ずしも十分に研究が行われているとは言えない．その一因として，日本語ツイートは一つ一つの文が短くまた文法的にも崩れていることも多く，形態素解析や構文解析などの精度が必ずしも高くないことが挙げられる．また，英語ツイートを対象とした感情分析では，Sentiment140 や NRC Affect Intensity Lexicon などの感情辞書を使用することで推定精度の向上を図ることが多いが，現状，日本語を対象とした感情辞書は必ずしも十分であるとは言えない．そのため，単純に文章中に現れる感情を表す単語とのマッチングを行うだけでは十分な推定精度を得ることが出来ないと考えられる．

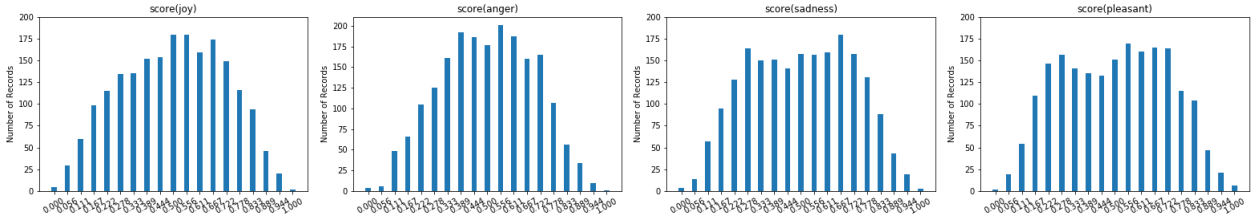
本研究では，近年研究が盛んな表現学習技術を用い，日本語ツイートを対象にその感情強度の推定問題に取り組む．具体的には，Best-Worst 尺度法 (Best-Worst Scaling) を用い，喜，怒，哀，楽の 4 感情に関する感情強度付き日本語ツイートデータセットの構築を行う．また作成したデータセットに対して種々の表現学習技術を適用し，感情辞書を利用しない感情強度推定モデルの構築とその実験的な評価を行う．

本論文の構成は以下の通りである．2 章で関連研究について述べる．3 章でデータセットの構築方法について述べた後，4 章で推定モデルの構築実験とその評価を行う．最後に 5 章でまとめを行い，今後の課題を述べる．

2　関連研究

WASSA-2017[1] は，初めて行われた感情強度推定タスクに関する国際コンペティションである．22 チームが参加しており，単語および文の分散表現と感情辞書による特徴抽出が最もよく使用された．ニューラルネッ

*連絡先：日本大学文理学部情報科学科
〒 156-8550 東京都世田谷区桜上水 3-25-40
E-mail: tozaki@chs.nihon-u.ac.jp



「喜」データ　「怒」データ　「哀」データ　「楽」データ

図 1: 各感情強度の分布

トワークが最も一般的に使用されている機械学習アーキテクチャであり，ツイートの表現学習や回帰関数への適合にも使用された．評価はピアソン相関係数が用いられ，ベンチマークシステムのピアソン相関係数は 0.66 となっている．

WASSA-2017 で最も優れたシステム Prayas[3] は，フィードフォワードニューラルネットワーク，マルチタスクディープラーニング，シーケンスモデリングの 3 種類のモデルのアンサンブル学習を行っている．提案モデルのピアソン相関係数は 0.747 であった．また YZU-NLP[4] は，感情辞書を使わない，BiLSTM と CNN を組み合わせたモデルを提案しており，提案したモデルのピアソン相関係数は 0.677 であった．YNU-HPCC[5] は，感情辞書を使用せず，CNN と LSTM を組み合わせることで，ツイート内の局地的な情報とツイート間の長距離依存関係を共に抽出し，感情強度を予測した．提案モデルのピアソン相関係数は 0.671 であった．

SemEval2018[2] でも，一つのタスクとして感情強度推定が採用されている．SemEval2018 の感情強度推定タスクで 1 位を獲得した SeerNet[6] は，4 種類の異なる事前学習済みのモデルに対して転移学習を行い，その後アンサンブル学習を行うものである．また，2 位を獲得した NTUA-SLP[7] は，多層自己アテンション付きの Bi-LSTM モデルを採用している．NTUA-SLP は，SemEval 2017 の Task4 のデータセットを用いてモデルの事前学習を行っており，その後転移学習を行っている．3 位となった PlusEmo2Vec[8] は，絵文字付きツイートで学習を行った 2 種類のモデルによる特徴量と，ハッシュタグを用いた感情に特化した単語分散表現の学習，エクスクラメーションマークやクエスチョンマークの数や大文字の単語の数などツイート特有の特徴量の 4 種類を用いたシステムとなっている．

一方，姫野ら [9] は，DeepMoji モデル [10] と呼ばれるモデルから得られる特徴量と EmoInt[11] による複数の感情辞書を用いた特徴抽出手法を組み合わせ，ツイートの感情強度を推定する手法を提案している．DeepMoji モデルは入力ツイートに適した絵文字を推定するモデルであり，絵文字が付いた 12 億ツイートを用いて学習されており，転移学習によって感情分析や皮肉分類などのタスクにおいても高い性能もつモデルである．

3　データセット

本章では，感情強度付き日本語ツイートデータセットの構築方法について述べる．

まず，1 名の専門家が手作業により，喜怒哀楽の各感情に対して各 1,000 件，合計 4,000 件の顔文字付きツイートを収集した．その後，収集した各ツイートから顔文字を取り除くことで同量の 4,000 件の（疑似的な）ツイートを準備した．これらの合計 8,000 件のツイートに対し，専門家 9 名によるアノテーションを行った．アノテーションの方法には，WASSA-2017 感情強度推定タスクにおけるデータセット作成でも採用された Best-Worst 尺度法（Best-Worst Scaling; BWS）を用いた．具体的には，同じ感情を持つ 4 ツイートを同時にアノテーターに提示し，最も感情強度が強いツイートと，最も感情強度が弱いツイートをそれぞれ選択させた．専門家によるアノテーション結果に基づき，各ツイート t の感情強度 $intensity(t)$ を式 (1) で算出した

$$intensity(t) = \left(\frac{most(t) - worst(t)}{n} + 1 \right) \times \frac{1}{2} \quad (1)$$

ここで $most(t)$ と $worst(t)$ は，それぞれツイート t が最も感情強度が強いと選択された回数と最も感情強度が弱いと選択された回数である．また n は，全アノテート作業中にツイート t が提示・比較された回数であり，今回の場合はアノテーター数の $n = 9$ である．$most(t)$，$worst(t)$ の値域は $[0, n]$ であり，$v = \frac{most(t) - worst(t)}{n}$ の値域は $[-1.0, 1.0]$ となるが，$intensity(t)$ の値域を $[0, 1]$ とするために v に対して線形変換を行った．

感情毎の感情強度の分布を図 1 に示す．図より「喜」感情に関しては正規分布に近い形をしているが，「哀」感情と「楽」感情では二峰性が確認できる．

表 1: モデルのハイパーパラメータ

	喜	怒	哀	楽	全
BiLSTM-CNN モデル					
単語の次元	200	100	300	300	200
BiGRU 隠れ層のサイズ	400	200	200	300	500
CNN フィルタサイズ	3	3	4	3	4
CNN フィルタ数	32	64	64	32	16
Pooling フィルタサイズ	2	2	2	2	3
全結合層のサイズ	32	64	16	16	16
ドロップアウト率	0.2	0.5	0.2	0.1	0.1
バッチサイズ	32	64	32	16	16
CNN-LSTM モデル					
単語の次元	100	200	200	200	300
CNN フィルタサイズ	3	3	4	3	3
CNN フィルタ数	64	64	64	16	32
Pooling フィルタサイズ	2	2	2	2	2
ドロップアウト率	0	0	0.1	0	0
GRU 隠れ層のサイズ	100	500	500	100	200
バッチサイズ	32	16	16	32	64

	喜	怒	哀	楽	全
Hybrid モデル					
単語の次元	100	200	100	300	300
文字の次元	50	50	100	50	50
BiGRU (word) 隠れ層のサイズ	400	400	400	100	200
CNN フィルタ数 (フィルタサイズ 2)	64	16	32	16	16
CNN フィルタ数 (フィルタサイズ 3)	16	64	16	64	16
CNN フィルタ数 (フィルタサイズ 4)	32	16	32	64	64
CNN フィルタ数 (フィルタサイズ 5)	64	16	32	32	16
GRU (chara) 隠れ層のサイズ	100	400	100	200	200
バッチサイズ	32	32	16	16	64

4 感情強度の推定実験

4.1 推定手法

今回，以下の 3 手法を用いて感情強度の推定モデルを構築した．

YZU-NLP[4] の BiLSTM-CNN モデル: 入力文を単語列とみなし，BiLSTM Layer，Convolution Layer，Max-Pooling Layer，Dense Layer, Output Layer の順に処理する．損失関数は平均二乗誤差 (mean squared error) を使用する．

YNU-HPCC[5] の CNN-LSTM モデル: 入力文を単語列とみなし，Convolution Layer，Max-Pooling Layer，LSTM Layer，Output Layer の順に処理する．損失関数は平均絶対誤差 (mean ablolute error) を使用する．

Dongyun[12] によるモデル: 単語レベルの表現と文字レベルの表現を組み合わせたモデルであり，単語レベルの入力は BiGRU Layer で処理される．また文字レベルの入力は，複数のフィルタサイズの Convolution Layer，GRU Layer の順に処理される．これら 2 つの出力は Highway Network を通り Output Layer で感情強度を出力する．このモデルは分類タスクを解くモデルであるため，L2 正則化を加えたクロスエントロピーを損失関数に用いていたが，今回の実験では回帰タスクでよく使用される平均二乗誤差 (mean squared error) で実装した．以降 Hybrid モデルと呼ぶ．

本実験では，これら 3 種類のモデルを用いて感情毎の感情強度推定に加え，4 種類の感情のデータセットを用いて感情の種類とその感情の強度推定を行った．

モデルの実装では，LSTM Layer の代わりに GRU Layer を使用した．またモデルの学習においては，最適化手法は adam を利用した．なお，エポック数を 30 としたが，5 エポック連続して評価データの損失値が改善しない場合，学習が収束したとみなし，学習を打ち切っている．

表 2: 感情強度の予測結果

	平均	喜	怒	哀	楽
BiLSTM-CNN	0.581	0.493	0.595	0.632	0.602
CNN-LSTM	0.562	0.500	0.623	0.662	0.464
Hybrid	0.768	0.743	0.701	0.808	0.820

4.2 感情毎の感情強度推定

本タスクでは，各感情 2,000 件のデータを，訓練データ 1,280 件，評価データ 320 件，テストデータ 400 件に分割して，学習・評価を行った．各モデルのハイパーパラメータは，ベイズ最適化によって評価データにおけるピアソン相関係数が最も高くなるパラメータを用いた．各モデルのハイパーパラメータを表 1 に，結果を図 2，表 2 に示す．なお図 2 は，各ツイート t に対する $intensity(t)$ の値 (x 軸) とその予測値 (y 軸) をプロットしたものである．

Hybrid モデルはすべての感情において他 2 つのモデルよりよい性能であり，4 感情のピアソン相関係数の平均は 0.768 であった．最も精度よく推定できた「楽」の感情のピアソン相関係数は 0.820 であった．最も推

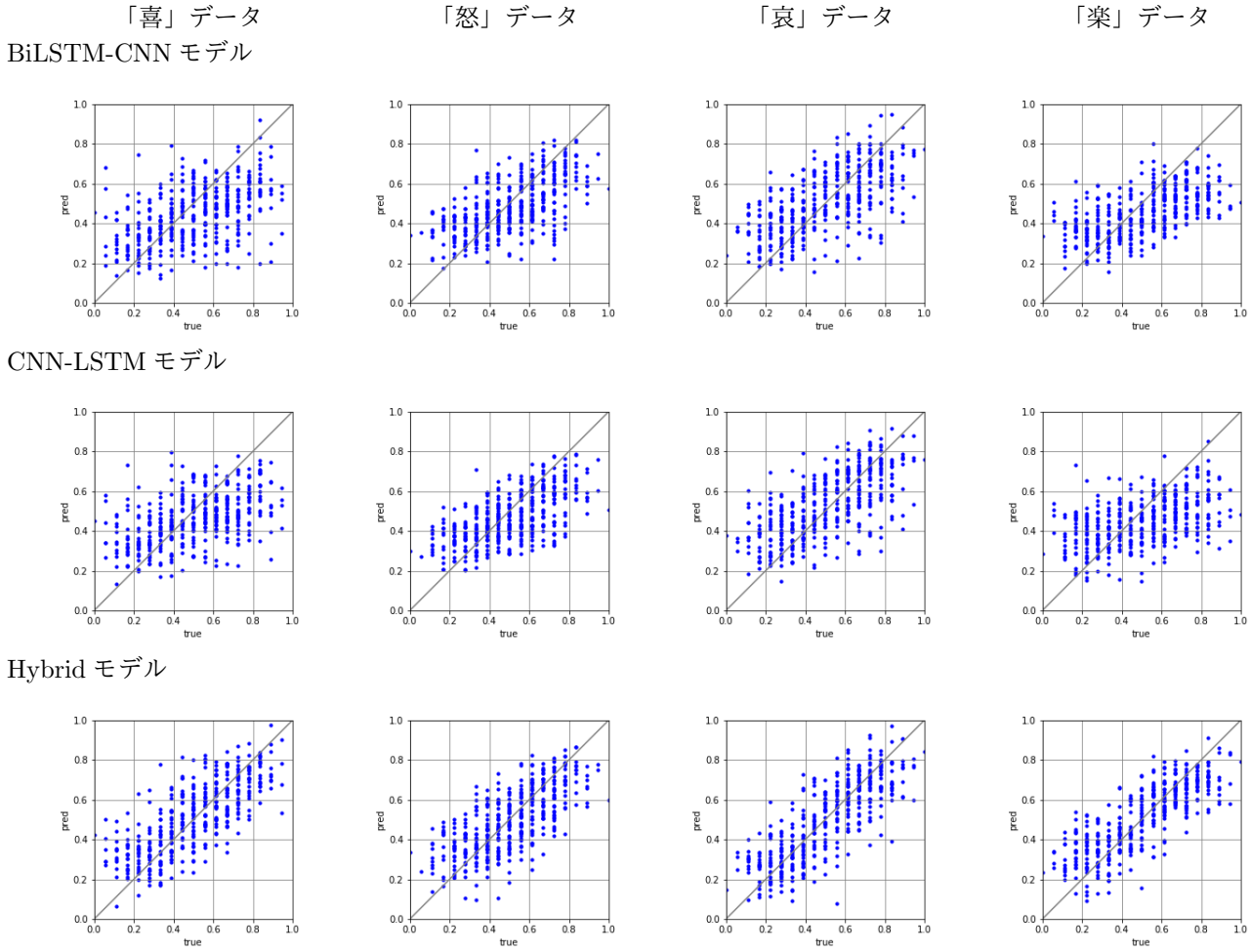


図 2: 感情強度の予測結果

定精度が低かった「怒」の感情のピアソン相関係数は 0.701 であったが，他 2 つのモデルより高いピアソン相関係数であった．BiLSTM-CNN モデルの 4 感情のピアソン相関係数の平均は 0.581 であった．最も精度よく推定できた「哀」の感情のピアソン相関係数は 0.632 であり，最も推定精度が低かった「喜」の感情のピアソン相関係数は 0.493 であった．CNN-LSTM モデルの 4 感情のピアソン相関係数の平均は 0.562 であった．最も精度よく推定できた「哀」の感情のピアソン相関係数は 0.662 であり，最も推定精度が低かった「楽」の感情のピアソン相関係数は 0.464 であった．

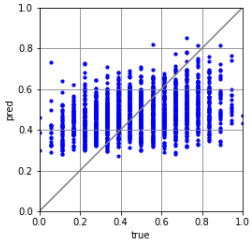
感情別に見ると，「哀」の感情が高い精度が得られ，「喜」の感情は低い推定結果が得られた．

4.3 感情種と感情強度の推定

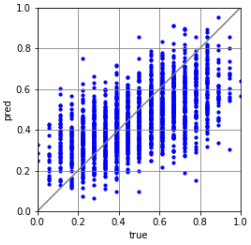
感情種とその強度の推定タスクでは，訓練データ 5,120 件（各感情 1,280 件ずつ），評価データ 1,280 件（各感情 320 件ずつ），テストデータ 1,600 件（各感情 400 件ずつ）に分割して，学習・評価を行った．本タスクでは，各モデルに感情分類を行う出力層を追加し，損失関数にはクロスエントロピーを使用した．ハイパーパラメータは，評価データにおけるピアソン相関係数と正確度 (accuracy) の和が最も高くなるパラメータを用いた．結果を図 3，表 3 に示す．また各モデルのハイパーパラメータは表 1 の”全”列に示す．

感情強度の推定精度が最も高いモデルは，Hybrid モデルであり，4 感情のピアソン相関係数の平均は 0.782 であった．CNN-LSTM モデルの 4 感情のピアソン相関係数の平均は 0.610，BiLSTM-CNN モデルは 0.266 となった．Hybrid モデルと CNN-LSTM モデルは，感情毎に感情強度を推定するよりも精度の向上が見られた．

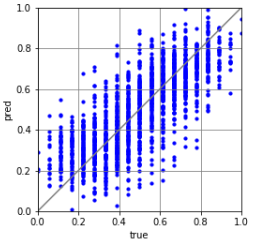
感情の分類精度が最も高いモデルは，CNN-LSTM モデルであり，4 感情の F 値の平均は 0.923 であった．Hybrid モデルは 4 感情の F 値の平均は 0.909，BiLSTM-CNN モデルは 0.896 となった．いずれのモデルにおいても，「哀」の感情の分類精度が最も低い結果となった．



BiLSTM-CNN モデル



CNN-LSTM モデル



Hybrid モデル

図 3: 感情強度ごとの予測値 (感情分類+感情強度)

表 3: 感情強度ごとの予測値 (感情分類+感情強度)

	precition	recall	f-measure	pearson
平均				
BiLSTM-CNN	0.906	0.908	0.896	0.266
CNN-LSTM	0.924	0.925	0.923	0.610
Hybrid	0.912	0.911	0.909	0.782
「喜」のデータの予測値				
BiLSTM-CNN	0.910	0.917	0.903	0.254
CNN-LSTM	0.941	0.917	0.927	0.613
Hybrid	0.930	0.912	0.919	0.800
「怒」のデータの予測値				
BiLSTM-CNN	0.900	0.936	0.907	0.450
CNN-LSTM	0.947	0.914	0.928	0.608
Hybrid	0.946	0.888	0.914	0.731
「哀」のデータの予測値				
BiLSTM-CNN	0.889	0.878	0.873	0.362
CNN-LSTM	0.904	0.928	0.914	0.620
Hybrid	0.863	0.908	0.884	0.743
「楽」のデータの予測値				
BiLSTM-CNN	0.924	0.902	0.903	-0.018
CNN-LSTM	0.904	0.944	0.922	0.599
Hybrid	0.907	0.934	0.919	0.853

5 まとめ

日本語ツイートの感情強度の推定を行うため,「喜」,「怒」,「哀」,「楽」の感情を持つツイートに対して，Best-Worst 尺度法を用いて，データセットの構築を行った．作成したデータセットに対して，既存の感情辞書を利用しない感情強度推定モデルの構築を行い，それぞれのモデルの評価を行った．また，感情毎の感情強度推定に加え，4 種類の感情のデータセットを用いて感情の種類とその感情強度の推定を行った．

感情毎の感情強度推定では，Hybrid モデルが最も良い性能であった．感情種と感情強度の推定では，感情分類精度は CNN-LSTM モデルが最も高い予測精度を出し，感情強度の推定精度では，Hybrid モデルが最も精度良く推定できる結果となった．

今後の課題として，教師なし学習でツイートの分散表現を獲得する手法と SVM などの機械学習手法による感情強度推定と，複数の分類器を用いたアンサンブル学習が挙げられる．

参考文献

[1] Saif M. Mohammad and Felipe Bravo-Marquez: WASSA-2017 Shared Task on Emotion Intensity, In Proceedings of the Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA), (2017)

[2] Saif M. Mohammad, Felipe Bravo-Marquez, Mohammad Salameh and Svetlana Kiritchenko: SemEval-2018 Task 1: Affect in Tweets, In Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval-2018), (2018)

[3] Pranav Goel, Devang Kulshreshtha, Prayas Jain and K.K. Shukla: Prayas at EmoInt 2017: An Ensemble of Deep Neural Architectures for Emotion Intensity Prediction in Tweets, In Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pages 58–65. (2017)

[4] Yuanye He, Liang-Chih Yu, K. Robert Lai and Weiyi Liu: YZU-NLP at EmoInt-2017: Determining Emotion Intensity Using a Bi-directional LSTM-CNN Model, In Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pages 238–242 (2017)

[5] You Zhang, Hang Yuan, Jin Wang and Xuejie Zhang YNU-HPCC at EmoInt-2017: Using a

CNN-LSTM Model for Sentiment Intensity Prediction, In Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pages 200–204 (2017)

[6] Venkatesh Duppada, Royal Jain and Sushant Hiray: SeerNet at SemEval-2018 Task 1: Domain Adaptation for Affect in Tweets, In Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval-2018), pages 18–23 (2018)

[7] Christos Baziotis, Nikos Athanasiou, Alexandra Chronopoulou, Athanasia Kolovou, Georgios Paraskevopoulos, Nikolaos Ellinas, Shrikanth Narayanan and Alexandros Potamianos: NTUA-SLP at SemEval-2018 Task 1: Predicting Affective Content in Tweets with Deep Attentive RNNs and Transfer Learning, In Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval-2018), pages 245–255 (2018)

[8] Ji Ho Park, Peng Xu, Pascale Fung: PlusEmo2Vec at SemEval-2018 Task 1: Exploiting emotion knowledge from emoji and #hashtags, In Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval-2018), pages 264–272 New Orleans, Louisiana, June 56 (2018)

[9] 姫野 晋之介, 青野 雅樹: 特徴量テンソルによる転移学習と感情辞書を用いた Twitter の感情強度推定, IEICE Technical Report DE2018-11 (2018-09) (2018)

[10] Bjarke Felbo, Alan Mislove, Anders Søgaard, Iyad Rahwan and Sune Lehmann : Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm, In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 1615–1625 (2017)

[11] Venkatesh Duppada and Sushant Hiray: Seernet at EmoInt-2017: Tweet Emotion Intensity Estimator, In Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pages 205–211 (2017)

[12] Dongyun Liang, Weiran Xu, Yinge Zhao: Combining Word-Level and Character-Level Representations for Relation Classification of Informal Text, In Proceedings of the 2nd Workshop on Representation Learning for NLP, pages 43-47 (2017)