# 主観感情と客観感情の強度推定のための日本語データセット

梶原 智之 †* Chenhui Chu‡* 武村 紀子 * 中島 悠太 * 長原 一 *
† 愛媛大学 ‡ 京都大学 * 大阪大学
{kajiwara, chu, takemura, n-yuta, nagahara}@ids.osaka-u.ac.jp

## 1 はじめに

感情分析は、対話システム [1] やソーシャルメディアマイニング [2] など多くの応用を持つ主要な自然言語処理タスクのひとつである。感情分析は活発に研究されており、テキストの感情極性の推定 [3] だけでなく、近年ではより詳細な感情の識別や感情強度の推定 [4] が試みられている。

先行研究では、Ekman の 6 感情（喜び・悲しみ・驚き・怒り・恐れ・嫌悪） [5] や Plutchik の 8 感情（喜び・悲しみ・期待・驚き・怒り・恐れ・嫌悪・信頼） [6] が代表的な基本感情として用いられている。既存の感情分析データセットには、テキストの書き手による主観的な感情強度のラベルを収集したもの [7] と、テキストの読み手による客観的な感情強度のラベルを収集したもの [8–13] が含まれる。多くの先行研究が客観的な感情強度ラベルを収集してきた[1]ため、これまでの感情分析の研究は客観的な感情強度の推定に焦点を当ててきた。

本研究では、テキストの書き手による主観的な感情強度ラベルとテキストの読み手による客観的な感情強度ラベルの両方を収集し、これらの差異について調査する。まず、クラウドソーシングを用いて雇用した 50 人の主観注釈者が、SNS 上での自身の過去の投稿に主観的な感情強度を付与した。そして、収集した全投稿に対して、それぞれ 3 人の客観注釈者が客観的な感情強度を付与した。合計で、17,000 件の投稿について、Plutchik の 8 感情 [6] の強度を、主観と客観の両方で 4 段階（無・弱・中・強）で付与した日本語の感情分析データセット[2]を構築した。

先述の通り、英語を中心に多くの感情分析データセット [7–13] が公開されているが、テキストの書き手が持つ感情と読み手が受ける感情の両方を収集したのは本研究が初である。日本語では、鍛治ら [14] や鈴木 [15] の研究があるが、これらは感情極性を対象としており、多様な感情を扱っていない。

**表 1** 感情強度ラベルの例（0：無、1：弱、2：中、3：強）
タイヤがパンクしてた。。いたずらの可能性が高いんだって。。

| | 喜び | 悲しみ | 期待 | 驚き | 怒り | 恐れ | 嫌悪 | 信頼 |
|---|---|---|---|---|---|---|---|---|
| 主観 | 0 | 3 | 0 | 1 | 3 | 0 | 0 | 0 |
| 客観 A | 0 | 3 | 0 | 3 | 1 | 2 | 1 | 0 |
| 客観 B | 0 | 2 | 0 | 2 | 0 | 0 | 0 | 0 |
| 客観 C | 0 | 2 | 0 | 2 | 0 | 1 | 1 | 0 |

収集した主観ラベルと客観ラベルを比較した結果、怒りや信頼の感情を中心に、読み手は書き手の感情を充分に検出できないことが明らかになった。表 1 に、書き手が強い悲しみと怒りの感情を持って書いた投稿の例を示す。読み手は書き手と同じく悲しみの感情を持っているが、怒りではなく驚きを感じており、書き手と読み手の感情は充分に一致しない。このように、書き手が強い怒りの感情を持って書いたテキストでも、その半数以上に読み手は怒りを全く感じないなど、全体的に読み手は書き手の感情を過小評価する傾向が見られた。また、BERT [16] による感情強度の推定実験の結果、客観ラベルよりも主観ラベルの推定が難しいことがわかった。テキストの書き手が持つ感情と読み手が受ける感情には大きなギャップが存在し、テキストの「読み手」である機械学習モデルにとっても、書き手の主観的な感情強度を推定することは難しいと言える。

## 2 感情分析データセットの構築

### 2.1 主観的感情強度ラベルの収集

クラウドソーシングサービスのランサーズ[3]を用いて、50 人の主観注釈者を雇用した。注釈者の内訳は、男性が 22 人、女性が 28 人、10 代が 2 人、20 代が 26 人、30 代が 18 人、40 歳以上が 4 人である。

1） EmoBank [10] は、テキストの書き手と読み手の両方の感情強度ラベルを収集したデータセットである。しかし、書き手とは異なるクラウドワーカが書き手の感情を推測しているため、書き手の主観的な感情を収集できているわけではない。

2） https://github.com/ids-cv/wrime

3） https://www.lancers.jp/

表 2　Quadratic Weighted Kappa による評価者間のアノテーションの一致率

| | 喜び | 悲しみ | 期待 | 驚き | 怒り | 恐れ | 嫌悪 | 信頼 | 全体 |
|---|---|---|---|---|---|---|---|---|---|
| 客観注釈者 A–客観注釈者 B | 0.697 | 0.607 | 0.594 | 0.342 | 0.627 | 0.359 | 0.527 | 0.203 | 0.547 |
| 客観注釈者 A–客観注釈者 C | 0.662 | 0.545 | 0.567 | 0.443 | 0.581 | 0.429 | 0.455 | 0.196 | 0.549 |
| 客観注釈者 B–客観注釈者 C | 0.700 | 0.597 | 0.632 | 0.415 | 0.630 | 0.476 | 0.512 | 0.295 | 0.585 |
| 主観注釈者–客観注釈者 A | 0.622 | 0.461 | 0.423 | 0.348 | 0.363 | 0.333 | 0.394 | 0.089 | 0.439 |
| 主観注釈者–客観注釈者 B | 0.633 | 0.526 | 0.432 | 0.339 | 0.386 | 0.361 | 0.442 | 0.153 | 0.465 |
| 主観注釈者–客観注釈者 C | 0.624 | 0.450 | 0.459 | 0.396 | 0.374 | 0.380 | 0.467 | 0.134 | 0.463 |
| 主観注釈者–客観注釈者の平均 | 0.683 | 0.536 | 0.498 | 0.441 | 0.401 | 0.433 | 0.514 | 0.132 | 0.515 |

主観注釈者は、SNS 上での自身の過去の投稿に対して、Plutchik の 8 感情 [6] の強度をそれぞれ 4 段階（無・弱・中・強）で付与した。ただし、テキストからの感情分析という目的のために、画像付きの投稿や URL 付きの投稿は対象外とした。各注釈者は 100 件から 500 件の投稿にラベルを付けており、合計で 17,000 件の投稿と主観的感情強度のラベルを収集した。投稿時期については特に制限を設けなかったが、結果として 2011 年 6 月から 2020 年 5 月までの 9 年の範囲の投稿が集まった。なお、投稿 1 件あたり 21.5 円の報酬を支払った。

収集したラベルの品質を評価するために、注釈者ごとに 30 件の投稿を無作為抽出した。評価者（大学院生 1 人）は、投稿内容と 8 感情の強度ラベルを見て、以下の基準で各投稿を 4 段階評価した。

- 3：付与されたラベルに完全に同意できる
- 2：付与されたラベルに概ね同意できる
- 1：付与されたラベルに同意しにくい
- 0：真面目にラベル付けしていると思えない

評価結果を注釈者ごとに平均すると、平均 2.1 点、最低 1.8 点、最高 2.5 点であった。なお、0 点の投稿は存在しなかった。平均 2 点を下回る注釈者は 5 人いたが、著しく低品質な注釈者は見られなかった。

### 2.2　客観的感情強度ラベルの収集

同じくランサーズを用いて、3 人の客観注釈者を雇用した。注釈者の内訳は、30 代の女性が 2 人と 40 代の女性が 1 人である。客観注釈者は、2.1 節で収集した 17,000 件の投稿に対して、主観注釈者と同じく、Plutchik の 8 感情 [6] の強度をそれぞれ 4 段階で付与した。ただし、主観注釈者がテキストの書き手である自身の感情を付与した一方で、客観注釈者はテキストの読み手である自身の感情を付与した。なお、投稿 1 件あたり 3.8 円の報酬を支払った。

収集したラベルの品質を評価するために、注釈者間の一致率として Quadratic Weighted Kappa[4] [17] を計算した。表 2 の上段に、客観注釈者間のアノテーションの一致率を示す。喜びの感情は $\kappa > 0.6$ の substantial agreement、信頼の感情は $\kappa < 0.4$ の fair agreement であるが、全体としては $0.5 < \kappa < 0.6$ の moderate agreement を確認できた。

### 2.3　主観注釈者の性格診断

本研究では、性格と感情の関係についても調査するために、主観注釈者に対して性格診断を実施した。性格 5 因子モデル [18] に基づく 60 項目の質問 [19] を通じて、調和性・外向性・情緒不安定性・開放性・誠実性の 5 種類の性格診断を実施した。この性格診断は、「陽気な」「素直な」などの 60 種類の性格形容語に対する自身の該当度を 7 段階評価で申告し、各項目の該当度の足し引きによって 5 種類の性格指標を計算するものである。4 節では、性格診断の結果を用いて感情分析モデルの改善を試みる。

## 3　データセットの分析

表 2 の下段に、主観注釈者と客観注釈者の間の感情強度の一致率を示す。客観注釈者間の感情強度の一致率に比べて、主観注釈者と客観注釈者の間の感情強度の一致率は全体的に低い。特に、怒りの感情において、客観注釈者間の一致率と主観注釈者–客観注釈者間の一致率に大きなギャップがある。なお、3 人の客観注釈者の感情強度を平均すると、主観注釈者との一致率が全体的に若干向上した。

表 3 に、主観の感情強度ラベルと客観の感情強度ラベルの混同行列を示す。例えば、喜びの感情において、主観注釈者が無とラベル付けした投稿のう

4）https://scikit-learn.org/stable/modules/generated/sklearn.metrics.cohen_kappa_score.html

表 3　主観ラベルと客観ラベルの混同行列（%）

| | 喜び | | | | 悲しみ | | | | 期待 | | | | 驚き | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 主観＼客観 | 無 | 弱 | 中 | 強 | 無 | 弱 | 中 | 強 | 無 | 弱 | 中 | 強 | 無 | 弱 | 中 | 強 |
| 無 | 91.7 | 3.1 | 4.0 | 1.2 | 90.2 | 4.7 | 4.0 | 1.1 | 84.1 | 6.7 | 6.0 | 3.1 | 80.9 | 7.8 | 7.8 | 3.5 |
| 弱 | 60.7 | 9.7 | 22.5 | 7.0 | 57.9 | 15.7 | 19.9 | 6.5 | 55.2 | 13.5 | 19.7 | 11.6 | 56.5 | 16.0 | 17.7 | 9.7 |
| 中 | 37.2 | 9.4 | 34.1 | 19.3 | 45.1 | 15.4 | 26.8 | 12.7 | 46.8 | 11.5 | 23.3 | 18.3 | 48.8 | 14.8 | 22.0 | 14.3 |
| 強 | 18.2 | 6.6 | 37.7 | 37.4 | 33.6 | 12.4 | 31.7 | 22.3 | 32.4 | 10.1 | 24.6 | 32.8 | 35.8 | 14.0 | 23.8 | 26.3 |

| | 怒り | | | | 恐れ | | | | 嫌悪 | | | | 信頼 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 主観＼客観 | 無 | 弱 | 中 | 強 | 無 | 弱 | 中 | 強 | 無 | 弱 | 中 | 強 | 無 | 弱 | 中 | 強 |
| 無 | 98.6 | 0.7 | 0.5 | 0.2 | 89.0 | 3.9 | 5.1 | 2.1 | 87.9 | 6.3 | 3.8 | 2.0 | 96.2 | 2.2 | 1.1 | 0.5 |
| 弱 | 87.5 | 5.2 | 4.8 | 2.5 | 70.3 | 8.7 | 14.0 | 7.0 | 62.2 | 16.1 | 13.1 | 8.6 | 92.6 | 4.2 | 2.2 | 1.1 |
| 中 | 77.4 | 7.2 | 9.6 | 5.8 | 57.5 | 9.3 | 19.7 | 13.5 | 49.1 | 14.5 | 18.9 | 17.5 | 86.3 | 6.5 | 5.1 | 2.0 |
| 強 | 58.6 | 6.7 | 15.2 | 19.5 | 44.2 | 6.9 | 22.8 | 26.1 | 34.7 | 11.4 | 21.3 | 32.6 | 81.5 | 7.2 | 6.4 | 4.8 |

ち、客観注釈者が無を付けた割合は 91.7%であり、客観注釈者が弱を付けた割合は 3.1%である。怒りの感情に注目すると、テキストの書き手である主観注釈者が強を付けた投稿の 58.6%に対して、テキストの読み手である客観注釈者は無を付けている。信頼の感情では更に顕著であり、主観注釈者が強を付けた投稿の 81.5%について客観注釈者は無を付けており、読み手が書き手の感情を充分に検出できないことがわかる。その他の感情についても、主観注釈者の感情強度が中以下の投稿については、客観注釈者によって無が付けられる割合が最も高い。また、主観注釈者の感情強度が無のときに客観注釈者の感情強度が弱以上になることは少なく、全体的に読み手は書き手の感情を過小評価する傾向がある。

## 4　感情強度推定の実験

2 節のデータセットを用いて、感情強度（無・弱・中・強）を推定する 4 値分類の評価実験を行う。

### 4.1　実験設定

データセットは、30 人 15,000 件の訓練用データ、10 人 1,000 件の検証用データ、10 人 1,000 件の評価用データに重複なく分割して実験する。単語分割には MeCab（IPADIC-2.7.0）[5] [20] を使用する。

感情分析モデルの性能は、平均絶対誤差によって評価する。主観注釈者が付与した感情強度ラベル（主観データ）を用いる評価と、3 人の客観注釈者が付与した感情強度ラベルの平均（客観データ）を用いる評価の両方を実施する。

感情分析モデルには、BERT [16] を用いる。事前訓練済みの BERT[6] を再訓練し、$\boldsymbol{y} = \mathrm{softmax}(\boldsymbol{h}\boldsymbol{W})$ として感情強度を推定する。ただし、$\boldsymbol{h}$ は BERT の [CLS] トークンから得られる素性ベクトルである。主観データにおける評価と客観データにおける評価の両方で、主観データで訓練した BERT（主観 BERT）と客観データで訓練した BERT（客観 BERT）の両方の性能を調査する。BERT の実装には Transformers[7] [21] を使用する。WholeWordMasking モデルを使用し、バッチサイズは 32、ドロップアウト率は 0.1、学習率は 2e-5、最適化は Adam [22] として、3 エポックの early-stopping を適用する。

主観データの評価では、主観 BERT において以下の 2 つの方法で主観注釈者の性格を考慮する。

- **w/ Pc**：$\boldsymbol{h}_c = [\boldsymbol{u}; \boldsymbol{v}]\boldsymbol{W}^c$ として、性格を考慮する。ただし、$\boldsymbol{u}$ は 5 次元の性格情報ベクトルを BERT と同じ 768 次元に線形変換した性格表現であり、$\boldsymbol{v}$ は BERT の [CLS] トークンから得られる 768 次元のテキスト表現である。感情強度の推定においては、$\boldsymbol{h}$ の代わりに $\boldsymbol{h}_c$ を用いる。
- **w/ Pa**：$\boldsymbol{h}_a = \mathrm{attention}(\boldsymbol{u}\boldsymbol{W}^Q, \boldsymbol{v}\boldsymbol{W}^K, \boldsymbol{v}\boldsymbol{W}^V)$ として、性格を考慮する。つまり、注意機構のクエリとして性格表現 $\boldsymbol{u}$、キーおよびバリューとしてテキスト表現 $\boldsymbol{v}$ を用いる。感情強度の推定においては、$\boldsymbol{h}$ の代わりに $\boldsymbol{h}_a$ を用いる。

5）https://taku910.github.io/mecab/

6）https://huggingface.co/cl-tohoku/bert-base-japanese-whole-word-masking

7）https://github.com/huggingface/transformers

表 4　感情強度推定の平均絶対誤差

| 主観データ | 喜び | 悲しみ | 期待 | 驚き | 怒り | 恐れ | 嫌悪 | 信頼 | 全体 |
|---|---|---|---|---|---|---|---|---|---|
| 最頻クラス | 0.896 | 0.713 | 0.907 | 0.684 | **0.218** | 0.344 | 0.435 | 0.429 | 0.578 |
| 主観 BERT | 0.734 | 0.666 | 0.899 | 0.684 | **0.218** | 0.344 | 0.443 | 0.432 | 0.553 |
| 主観 BERT w/ Pc | 0.784 | 0.698 | 0.870 | 0.659 | **0.218** | **0.343** | 0.457 | 0.429 | 0.557 |
| 主観 BERT w/ Pa | 0.740 | 0.665 | 0.850 | 0.665 | **0.218** | 0.351 | 0.441 | 0.429 | 0.545 |
| 客観 BERT | **0.674** | **0.623** | **0.789** | **0.634** | **0.218** | 0.356 | **0.432** | **0.427** | **0.519** |
| 客観注釈者 A | 0.545 | 0.544 | **0.713** | 0.686 | 0.211 | 0.523 | 0.522 | 0.428 | 0.522 |
| 客観注釈者 B | **0.521** | **0.520** | 0.720 | **0.571** | 0.201 | **0.347** | **0.375** | **0.426** | **0.460** |
| 客観注釈者 C | 0.526 | 0.533 | 0.738 | 0.694 | **0.200** | 0.610 | 0.520 | 0.432 | 0.532 |
| 客観データ | 喜び | 悲しみ | 期待 | 驚き | 怒り | 恐れ | 嫌悪 | 信頼 | 全体 |
| 最頻クラス | 0.595 | 0.459 | 0.713 | 0.518 | **0.044** | 0.420 | 0.383 | 0.026 | 0.395 |
| 主観 BERT | 0.489 | 0.446 | 0.685 | 0.518 | **0.044** | 0.414 | 0.381 | 0.039 | 0.377 |
| 客観 BERT | **0.403** | **0.411** | **0.475** | **0.442** | **0.044** | **0.386** | **0.348** | **0.024** | **0.317** |
| 客観注釈者 A | **0.222** | 0.246 | 0.277 | **0.276** | **0.035** | **0.233** | **0.226** | 0.053 | **0.196** |
| 客観注釈者 B | 0.224 | 0.264 | **0.268** | 0.349 | 0.045 | 0.277 | 0.269 | 0.027 | 0.215 |
| 客観注釈者 C | 0.237 | **0.241** | 0.332 | 0.360 | 0.046 | 0.346 | 0.310 | **0.021** | 0.237 |

参考のために、最頻クラスのベースラインも評価する。これは、感情ごとに最も高頻度なラベルを常に出力するベースラインである。本データセットではいずれの感情においても無のラベルが最高頻度であるため、実際には常に無のラベルを出力する。

### 4.2　実験結果

実験結果を表 4 に示す。下段の客観データにおける評価よりも上段の主観データにおける評価の方が全体的に平均絶対誤差が大きく、書き手の感情強度の推定がより難しいことがわかる。これまでの議論で、書き手の感情を読み手が推定するのは難しいと述べたが、同じく「読み手」である機械学習モデルにとっても書き手の感情推定は難しいと言える。

上段の主観データにおける評価では、主観データで訓練した主観 BERT ではなく客観データで訓練した客観 BERT が最高性能を達成するという想定外の結果となった。先述のように、テキストの書き手の感情（主観データ）の推定は難しいため、単純な訓練では充分な性能が得られなかったものと考える。

そこで、書き手の性格情報を用いて訓練を補助する 2 手法を検討した。実験の結果、性格表現とテキスト表現を単純に結合する主観 BERT w/ Pc は有効ではなかったが、性格表現とテキスト表現を重み付きで考慮する主観 BERT w/ Pa は単純な主観 BERT よりも低い平均絶対誤差（0.553 → 0.545）を達成した。しかし、性格情報を用いても、客観 BERT の性能には及ばなかった。

参考のために、客観注釈者の性能も掲載する。客観 BERT の性能は、客観注釈者 A および C を超えており、テキストの「読み手」としての主観感情の推定においては人間に匹敵する水準にあると言える。

下段の客観データにおける評価でも、主観データにおける評価と同じく、BERT は最頻クラスのベースラインよりも低い平均絶対誤差を達成し、客観 BERT が最高性能を達成した。各客観注釈者と比較すると、客観 BERT の性能は、客観感情の推定においては改良の余地が残っていることがわかる。

## 5　おわりに

本研究では、日本語の感情分析のために 17,000 件のデータセットを構築し、公開[2)] した。Plutchik の 8 感情に基づき、50 人の注釈者が自身の過去の SNS の投稿に主観的な感情強度をラベル付けし、さらに 3 人の注釈者が客観的な感情強度を付与した。

書き手の主観的な感情強度と読み手の客観的な感情強度を比較したところ、怒りや信頼の感情を中心に、読み手は書き手の感情を充分に検出できず、過小評価する傾向が見られた。また、同じくテキストの「読み手」である機械学習モデルにとっても、書き手の主観的な感情強度の推定は難しい。今後は、書き手の性格情報や過去の投稿履歴を考慮し、より高精度に主観感情を推定するモデルを開発したい。

## 謝辞

本研究は文部科学省による Society 5.0 実現化研究拠点支援事業の支援を受けたものである。

# 参考文献


[1] 徳久良子, 乾健太郎, 松本裕治. Web から獲得した感情生起要因コーパスに基づく感情推定. 情報処理学会論文誌, Vol. 50, No. 4, pp. 1365–1374, 2009.

[2] Stefan Stieglitz and Linh Dang-Xuan. Emotions and Information Diffusion in Social Media — Sentiment of Microblogs and Sharing Behavior. *Journal of Management Information Systems*, Vol. 29, No. 4, pp. 217–248, 2013.

[3] Richard Socher, Alex Perelygin, Jean Y. Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 1631–1642, 2013.

[4] Laura Ana Maria Bostan and Roman Klinger. An Analysis of Annotated Corpora for Emotion Classification in Text. In *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 2104–2119, 2018.

[5] Paul Ekman. An Argument for Basic Emotions. *Cognition and Emotion*, Vol. 6, No. 3–4, pp. 169–200, 1992.

[6] Robert Plutchik. A General Psychoevolutionary Theory of Emotion. *Theories of Emotion*, Vol. 1, pp. 3–31, 1980.

[7] Klaus R. Scherer and Harald G. Wallbott. Evidence for Universality and Cultural Variation of Differential Emotion Response Patterning. *Journal of Personality and Social Psychology*, Vol. 66, No. 2, pp. 310–328, 1994.

[8] Saima Aman and Stan Szpakowicz. Identifying Expressions of Emotion in Text. In *International Conference on Text, Speech and Dialogue*, pp. 196–205, 2007.

[9] Carlo Strapparava and Rada Mihalcea. SemEval-2007 Task 14: Affective Text. In *Proceedings of the Fourth International Workshop on Semantic Evaluations*, pp. 70–74, 2007.

[10] Sven Buechel and Udo Hahn. EmoBank: Studying the Impact of Annotation Perspective and Representation Format on Dimensional Emotion Analysis. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 578–585, 2017.

[11] Saif Mohammad and Felipe Bravo-Marquez. Emotion Intensities in Tweets. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics*, pp. 65–77, 2017.

[12] Saif Mohammad and Svetlana Kiritchenko. Understanding Emotions: A Dataset of Tweets to Study Interactions between Affect Categories. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, pp. 198–209, 2018.

[13] Laura Ana Maria Bostan, Evgeny Kim, and Roman Klinger. GoodNewsEveryone: A Corpus of News Headlines Annotated with Emotions, Semantic Roles, and Reader Perception. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pp. 1554–1566, 2020.

[14] 鍛治伸裕, 喜連川優. HTML 文書集合からの評価文の自動収集. 自然言語処理, Vol. 15, No. 3, pp. 77–90, 2008.

[15] Yu Suzuki. Filtering Method for Twitter Streaming Data Using Human-in-the-Loop Machine Learning. *Journal of Information Processing*, Vol. 27, pp. 404–410, 2019.

[16] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4171–4186, 2019.

[17] Jacob Cohen. Weighted Kappa: Nominal Scale Agreement Provision for Scaled Disagreement or Partial Credit. *Psychological Bulletin*, Vol. 70, No. 4, pp. 213–220, 1968.

[18] Lewis R. Goldberg. The Development of Markers for the Big-Five Factor Structure. *Psychological Assessment*, Vol. 4, No. 1, pp. 26–42, 1992.

[19] 齊藤崇子, 中村知靖, 遠藤利彦, 横山まどか. 性格特性用語を用いた Big Five 尺度の標準化. 九州大学心理学研究, Vol. 2, pp. 135–144, 2001.

[20] Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. Applying Conditional Random Fields to Japanese Morphological Analysis. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pp. 230–237, 2004.

[21] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, 2020.

[22] Diederik P. Kingma and Jimmy Lei Ba. Adam: A Method for Stochastic Optimization. In *Proceedings of the 3rd International Conference on Learning Representations*, 2015.