

アマゾン商品レビューの感情分析と評価スコアの相関性の 検証

高山 なつ樹^{1,a)} Mario Köppen^{2,b)} 大北 剛^{2,c)}

概要：インターネットにおける商品レビューを感情分析をした場合のネガポジ判定は、商品の評価スコアに比例するのではないかという仮説を立てた。これを検証するため、商品レビューと評価の両方のデータを合わせ持つアマゾンの商品レビューの評価を2値に分類して、商品レビューに対する感情分析における極性とマッチするかどうかを検証した。4種類の学習器をベースとした感情分析器を用いて、クラス不均衡問題を考慮しながら、感情分析を行なった。結果は、ロジスティック回帰を用いた感情分析を用いると、F1 指標で 93 パーセントとなり、上記の仮説が正しいことを実験的に確かめた。

Verification of Correlation between Sentiment Analysis and Rating Score of Amazon Product Review

1. はじめに

感情とは、特定の主題やトピックに対して個人がどのように感じているのかを表すために使われる主観的な表現を指す。感情分析とは、テキストから意見や感情あるいは態度を理解するために利用される比較的新しい分野のデータマイニングである。感情分析により、特定の商品等に対する一般消費者の意見を定量化することができる。本稿では、インターネットにおける商品レビューを感情分析をした場合のネガポジ判定は、商品の評価スコアに比例するのではないかという仮説を立てた。この仮説の検証のために、商品レビューが、テキストと1から5までのレート評価から構成されるアマゾンの商品レビューを採用した。

2. 使用したデータセット

データ解析コンペティションのプラットフォーム Kaggle で公開されているデータセット”Amazon Reviews: Unlocked Mobile Phones”を学習用, ”Consumer Reviews of

Amazon Products”を評価実験用として使用した。

□ ”Amazon Reviews: Unlocked Mobile Phones”

PromptCloud より提供される。アマゾンで販売されているロック解除された携帯電話の 413840 件のレビューのリストである。レビュー、評価、価格等のカラムで構成される。

□ ”Consumer Reviews of Amazon Products”

Datafiniti の製品データベースより提供される。アマゾンで販売されている Kindle, Fire TV Stick 等に関する 34660 件のレビューのリストである。各製品の基本的な製品情報、レビュー、評価等のカラムで構成される。

アマゾンの前者のデータをトレーニング集合とテスト集合(1つ目のテスト集合とする)と考え、分析器の構築に使用する。一方、後者のデータにおいてほとんどのレビューに対してはレート評価があるが、レート評価のないものが 33 件存在する。そこで、この 33 件のレビューを2つ目のテスト集合と考え、人手でレート評価を与えたのち構築した分析器をテストした。

3. 我々の感情分析手法

我々の目的は、インターネットにおける商品レビューを感情分析をした場合のネガポジ判定は、商品の評価スコアに比例するのではないかという仮説を検証することである。

¹ 九州工業大学 大学院 情報工学府
Graduate School of Computer Science and Systems Engineering, Kyushu Institute of Technology

² 九州工業大学
Kyushu Institute of Technology

a) kouyama.natsuki731@mail.kyutech.jp

b) mkoepen@ieee.org

c) tsuyoshi.okita@gmail.com

表 1 データセットの詳細

データセット (略称)	レビューの数	レビューの平均文字長	使用目的
"Amazon Reviews: Unlocked Mobile Phones" (アマゾン A)	413840	264.6	トレーニングとテスト (分析器の構築)
"Consumer Reviews of Amazon Products" (アマゾン B)	34660	263.0	33 件の無レート評価レビューの抽出, テスト

そこで、この目的を達成するために、以下のような形で感情分析器を構築した。まず機械学習可能なデータとするため、ラベルを整える作業と特徴抽出を行う。具体的にはレート評価を 2 値のポジティブとネガティブというラベルに変換する作業を行い、一方、レビューのテキストに対して正規化、小文字変換などの前処理をした後、tf-idf 特徴抽出を行った。以上により機械学習可能な形となり入出力ペア (X, Y) が揃う。そこでこのデータに対して感情分析器を構築する。感情分析器としてはランダムフォレスト、多項分布ナイーブベイズ、ベルヌーイナイーブベイズ、ロジスティック回帰という 4 種類を用いる。このようにして構築した感情分析器を用いて、3 種類の検証を行う。

つまり、以下のような分析フローを行う。

- (1) データセットの用意：学習器作成用と評価実験用に 2 つの異なるデータセットを用意する。
- (2) データセットの前処理：レート評価に基づきネガポジの 2 値ラベリング、さらにレビューテキストのクリーニング処理や単語の正規化を行う。
- (3) 特徴抽出：ネガポジラベリングしたレビューのテキストにおける頻出単語を調べ、tf-idf 特徴を抽出して重み付けする。
- (4) 学習器の作成：4 種類の学習器をベースとした感情分析器を作成する。
- (5) 学習器の評価：作成した感情分析器を各スコアにより評価する。
- (6) 評価実験：自作のレビューテキストを分析器に与えてテストし、さらに仮説の検証を行う。

以下これらの各手順について詳しく述べる。

3.1 ネガポジラベリング

アマゾン A のデータセットについて、1 から 5 で表されるレート評価に従って、レビューをポジティブとネガティブの 2 値でラベリングしたいと考えた。まず、図 1 にレート評価の分布を示す。この図が示すように、本データセットには各レートの評価数に偏りが見られるため、単純にレート評価 1,2,3 をネガティブ、4,5 をポジティブとするのは不適切であると考えた。従ってレート評価 1,2 をネガティブ、3,4 をポジティブとラベリングし、5 は使用しない。このようにネガポジラベリングした後のそれぞれのレビュー数の分布を図 2 に示す。

3.2 レートのないレビュー

アマゾン B のデータセットについて、約 3 万件のレビューの内 33 件にレート評価がなかった。これらは、実験時に感情分析器でレビューテキストの分析をするために使用する。

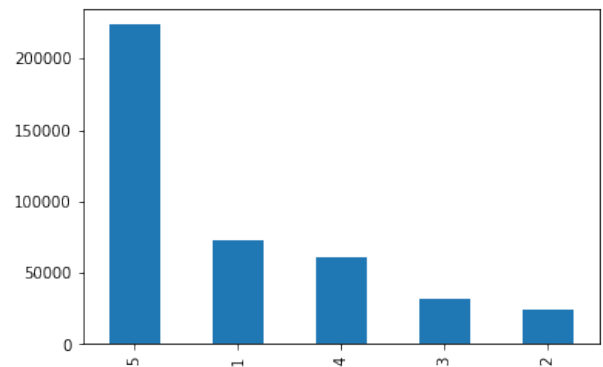


図 1 レート評価の分布

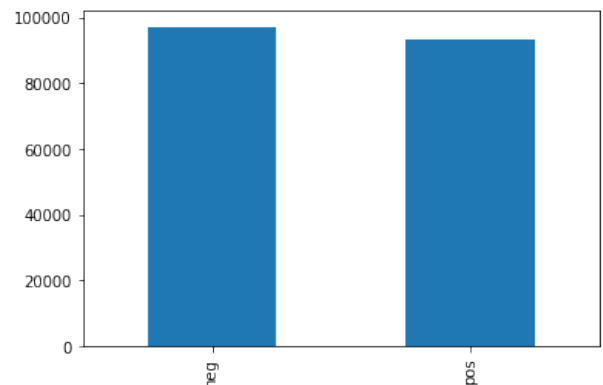


図 2 ラベリング後のネガポジ分布

3.3 テキストの前処理

今後の分析をより正確に行うため、また計算量やメモリ使用量の観点から見ても重要な処理となるテキストのクリーニング処理を行う。本稿におけるレビューテキストは全て英文である。従って、大文字を全て小文字に変換してテキストの正規化を行い、句読点を除去した。表 2 に、クリーニング処理前後のレビューテキストの一例を示す。

3.4 tf-idf 特徴抽出

tf-idf は文書に含まれる単語の重要度を評価する尺度の一つである。頻度の低いレアな単語が(ローカルな)文書中に何度も出て来る場合、この単語の重要度を上げることにより、文書分類を行う際のこの単語の重みを重くすると

表 2 クリーニング処理前後のレビューテキスト

	クリーニング処理前	クリーニング処理後
例 1	nice phone, nice up grade from my pantach revu...	nice phone nice up grade from my pantach revue...
例 2	It works good but it goes slow sometimes but i...	it works good but it goes slow sometimes but i...
例 3	Great phone to replace my lost phone. The only...	great phone to replace my lost phone the only ...

いう手法である。

ローカルな各文書における単語の出現頻度 (term frequency) を

$$tf(t,d) = \frac{\text{文書 A における単語 X の出現頻度}}{\text{文書 A における全単語の出現頻度の和}}$$

と定義し、逆文書頻度 (inverse document frequency) を

$$idf = \log\left(\frac{\text{全文書数}}{\text{単語 X を含む文書数}}\right)$$

と定義する。これらを用いて、tf-idf が

$$tf-idf = tf \times idf$$

$$= \text{単語の出現頻度} \times \text{各単語のレア度}$$

と定義される。

レビュー全体の語はアマゾン A のデータ集合においては、413840 語存在するが、語彙としては 83969 語存在した。1 つのレビューは通常複数の文を含む形で、文章を分けてはいない。文章でそれぞれの語彙の頻度をカウントし、さらにこれらそれぞれの文章の語がそれぞれのレビューにおける tf-idf 特徴を抽出した。このことにより、レビュー毎に、83969 個の bag-of-words のリストを得る。ID0 の語は常にこのリストの先頭にあり、それぞれの文におけるこの ID0 の単語の tf-idf 特徴がこの項目にあたる。このリストはレビューの数だけ存在するため、この場合 190235 のレビューの数だけ存在した。つまり、190235 行の 83969 列の行列を考えることになる。つまり、この行列の行はレビュー ID を表わし、列は語彙 ID を表わし、行列の中身は tf-idf 特徴を示す。

3.5 感情分析器の構築

このようにして抽出した語彙毎の tf-idf 特徴に関するレビューを $\mathcal{X} (= [x_1, \dots, x_N] (x_i \in \mathcal{R}))$ として、それぞれのレーティングをネガポジの 2 値ラベル $\mathcal{Y} (\in \{0, 1\})$ として、学習器を構成した。つまり、学習器は $f: \mathcal{X} \rightarrow \mathcal{Y}$ となるような関数を学習する。

このような学習器はさまざまなものが考えられるため、ここではランダムフォレスト、多項分布ナイーブベイズ、ベルヌーイナイーブベイズ、ロジスティック回帰の 4 種類の学習器をベースとするものを考えた。

それぞれの感情分析器の設定は以下の通りである。ランダムフォレストの設定は、max_depth=30, n_estimators=30, random_state=42 とした。多項分布ナイーブベイズは

$\alpha = 1.0$ 、同じくベルヌーイナイーブベイズも $\alpha = 1.0$ とした。ロジスティック回帰は学習器、正則化の組み合わせとして (lbfg/liblinear/newton-cg/sag/saga, L1/L2/elasticnet) などの選択肢が可能であり、一方、正則化パラメータ C として [1, 10, 100, 1000, 10000, 100000] という値域を設定し、グリッドサーチを行なって性能の良いパラメータ探索を行った。すべての条件として、クラス不均衡を校正する形で走らせた。この結果、正則化パラメータ C が 100 の場合に saga を用いた L2 最適化が最も性能が良いという結果を得た。この場合、正答率は 0.9396 であった。

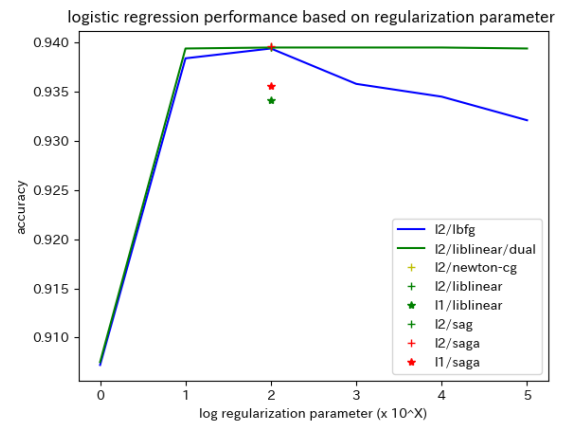


図 3 ロジスティック回帰におけるグリッドサーチによる性能の良い学習器の探索

3.6 構築した感情分析を用いた検証

このようにしてした構築した感情分析器を用いて以下の検証を行う。

- (1) 商品レビューに対する感情分析における極性とマッチするかどうかを検証する
- (2) 手書きレビューを与えて精度を確かめる
- (3) レート評価のないレビューの極性を分類する

これらの検証を構築した 4 種類の学習器それぞれに対して行なった。

3.7 感情分析の説明可能性の検証とネガポジ語彙の抽出

ロジスティック回帰を用いて構築した感情分析器を用いて、SHAP 値 [1] を手掛かりにそれぞれのレビューにおいてネガティブかポジティブの判断をした根拠を示すことができる。つまり、レビュー毎にどの語を理由としてネガ

ティブかポジティブの判断をしたかを検証する。

一方、この方法をグローバルにレビュー全体の 83969 件に対して行い、SHAP 値の高い順に語彙を抽出することにより、ネガポジに影響力の高い語が抽出できるはずである。したがって、これらをアマゾンの商品に関する感情語彙とする。

4. 結果

4.1 各スコアと ROC 曲線

表 3 に 4 種類の分析器をベースにした感情分析器のスコアを、図 4 にはその ROC 曲線を示す。

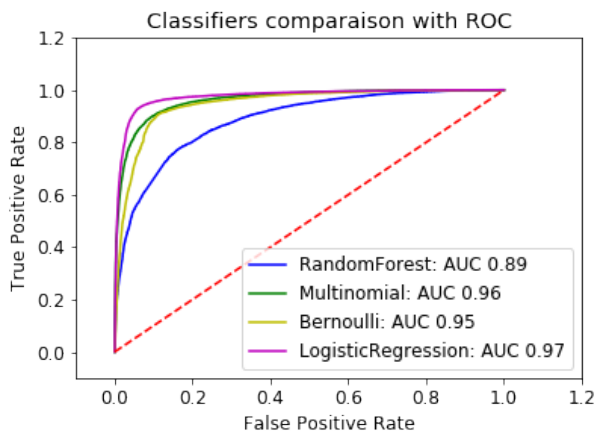


図 4 4 種類の感情分析器の ROC 曲線

4.2 手書きレビューを用いたテスト

最も精度の良いロジスティック回帰をベースとした感情分析器について、手書きのレビューテキストを与えて感情の極性を予測した。表 4 にポジティブなレビューを、表 5 にネガティブなレビューをそれぞれ 3 つずつ与えた場合の結果を示す。

4.3 レートのないレビューに対するテスト

アマゾン B のデータセットの 33 件のレート評価のないレビューについて、ロジスティック回帰を用いた感情分析器により感情極性を予測した。また、実際に筆者がレビューテキストを読み、ネガポジのラベリング後、モデルが予測した極性とマッチするか調べた。結果として 33 件中 4 件が誤答であり、正答率 88% を得た。

4.4 感情分析の説明可能性から得られるネガポジを示す語

ロジスティック回帰を用いて構築した感情分析器を用いて、SHAP 値 [1] を手掛かりに判断をした根拠を示すことができる。

これをさらに押しすすめることにより、アマゾン商品のレビューにおける感情語彙を作ることができた。つまり、

SHAP 値が高い順に語彙を抽出すると、感情語彙としてネガポジに影響力の高い語が抽出できるはずである。図 6 と図 7 はそれぞれそのような感情語彙のリストを示す。図において上位にある特徴量は極性の予測により影響力が大きいことを示す。極性は負が正なので、横軸はこれらのいずれかを取り、例えば、一番上の great は右側へ青から赤線が伸びており、これは great がレビューに含まれると正の極性への影響が大きいことを示す。逆に not があると左側へ青から赤線が伸びており、負の極性へ導く影響が大きいことを示す。

このアマゾン商品のレビューにおける商品は電話であることから、good phone があると、その評価はかなり良いということがわかる。また、works well/not work/love phone/not recommended/not good/nice phone/not worth/not happy など 2 グラムとしてロジスティック回帰を走らせたことで、これらの語の影響を結果として含んでいることは、1 グラムでなく、2 グラムを用いたほうが効果的であることを示している。

また、not buy/junk/not happy/don buy/garbage などは青の点線を示している。これらは負 (neg) より X 軸のかなり左に表われており、これらの語が含まれたことで、特に負の極性に影響することを示している。

感情分析は本来文章の感情がポジティブか、ネガティブかという極性を求めるが、本論文においては、これらのラベルを商品の評価と置き換えて学習をした。このことから、感情分析ではあるが、上記で得られたネガポジを示す語は、感情の分析を示す語と商品の良し悪しに関係する語を混合した結果となった。これは予想通りである。Hamilton et al.[2] はドメインに特定した感情語彙を求める手法を考案しており、そこで誘導された語を語彙のリストとして提供している。たとえば、ACTrade というのは reddit でコンピュータ関連グッズをやり取りするためのリストであり、アマゾンの電話とは若干対象が違うとは言え、同等の語彙が上位に上がってくると予想される。そこでの負のトップ 20 は bad/suck/sucks/awkward/dumb/stupid/pitfall/test/450k/messed/failed/exact/safe/power/pricing/cut/suddenly/in-game/randomly/slow である。一方、正のトップ 20 は wonderful/grateful/wow/thankyou/appreciated/appreciate/generosity/much/kindness/thanks/brilliant/greatly/ooh/thank/yw/ pleasure/definitely/common/patient/favor である。これらはコミュニティポストという形式であるため、これらを反映した語も含むためもあるが、アマゾンの電話のレビューとはかなり異なる語彙であることがわかる。

表 3 4 種類の分析器のスコア

	ランダムフォレスト	多項分布 NB	ベルヌーイ NB	ロジスティック回帰
pos-precision	0.79	0.88	0.79	0.94
neg-precision	0.83	0.92	0.95	0.93
pos-recall	0.85	0.92	0.93	0.93
neg-recall	0.76	0.87	0.73	0.93
accuracy	0.80	0.90	0.84	0.93

表 4 ロジスティック回帰による手書きレビューの感情予測（ポジティブの例）

レビューの内容	予測した感情	ネガティブ確率	ポジティブ確率
"The product was good and easy to use"	POS	0.000010	0.999990
"I will recommend to my friend"	POS	0.000000	1.000000
"not bad"	POS	0.038882	0.961118

5. 考察

5.1 語彙から極性を判断するアプローチ

本論文で扱った形の分析は、語彙から極性を判定するものではない。古典的な感情分析においては、語彙をベースに極性を判定していた（バレンス法）。図 5 はそのような方法に基づき描いた極性判定の分布を示す。それぞれの語がどのような極性をもつかのデータベースとして定義された場合に、バレンスは、これらの語に基づいて、文章の極性がこれらの平均として求めたものである。ルールベースの手法である。

商品プロダクトのレビューに関して、通常の感情分析の語彙に基いた判定を行うと、大半がポジティブでもネガティブでもなくゼロ付近に固まっているが、全体としてネガティブな文書の数より、ポジティブな文書の数が多いことがわかる。

参考文献

- [1] Lundberg, Scott M and Lee, Su-In, A Unified Approach to Interpreting Model Predictions, Advances in Neural Information Processing Systems 30, pp.4765–4774, 2017.
- [2] William L. Hamilton, Kevin Clark, Jure Leskovec, and Dan Jurafsky. Inducing Domain-Specific Sentiment Lexicons from Unlabeled Corpora. ArXiv preprint (arxiv:1606.02820). 2016.

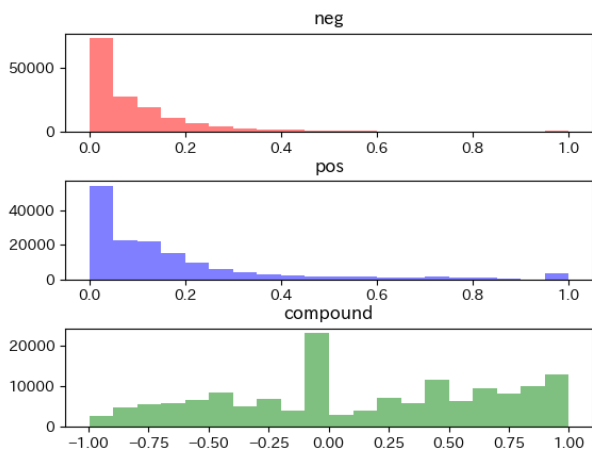


図 5 バレンス。

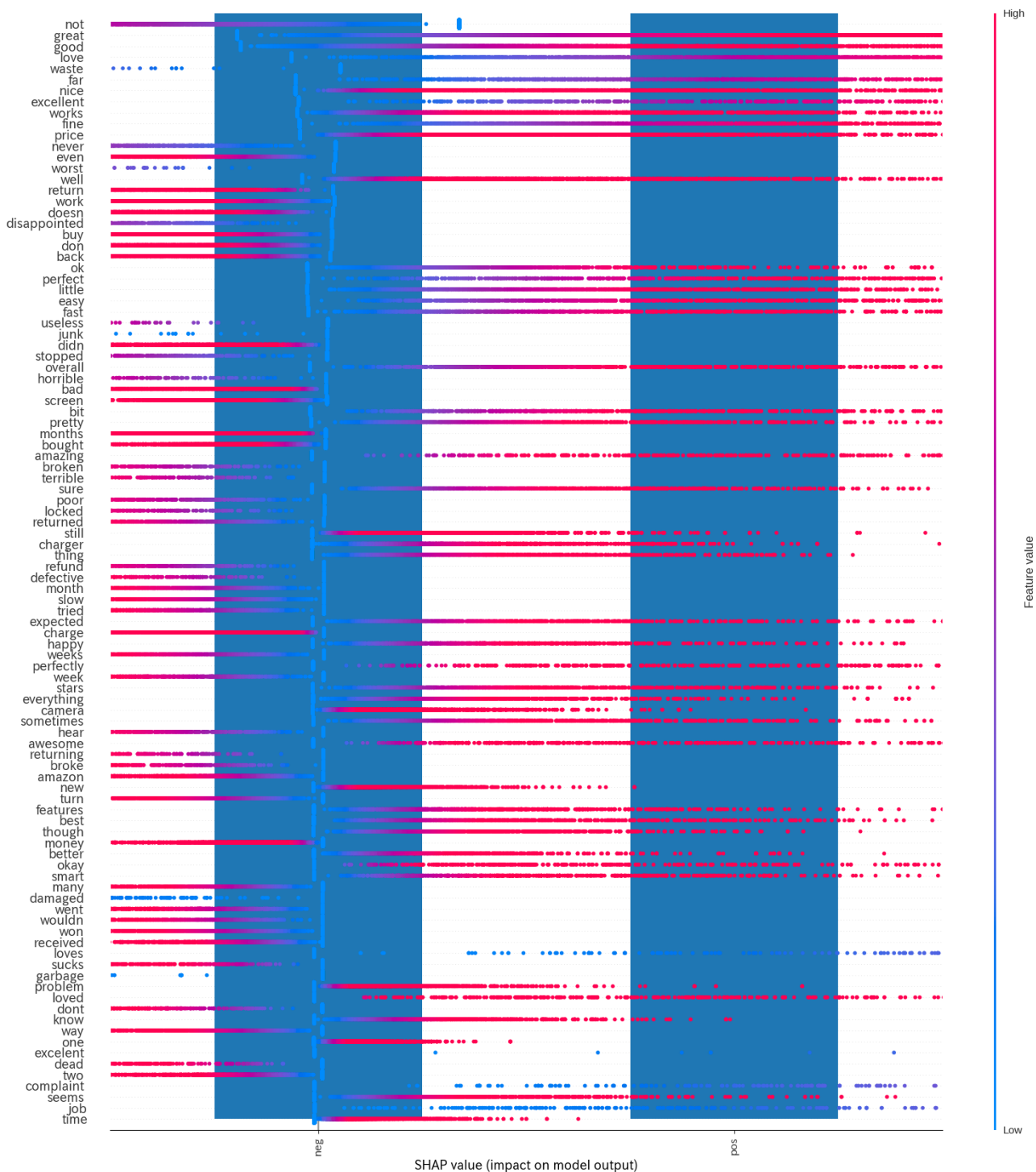


図 6 1gram での SHAP 値の高い語のリスト.

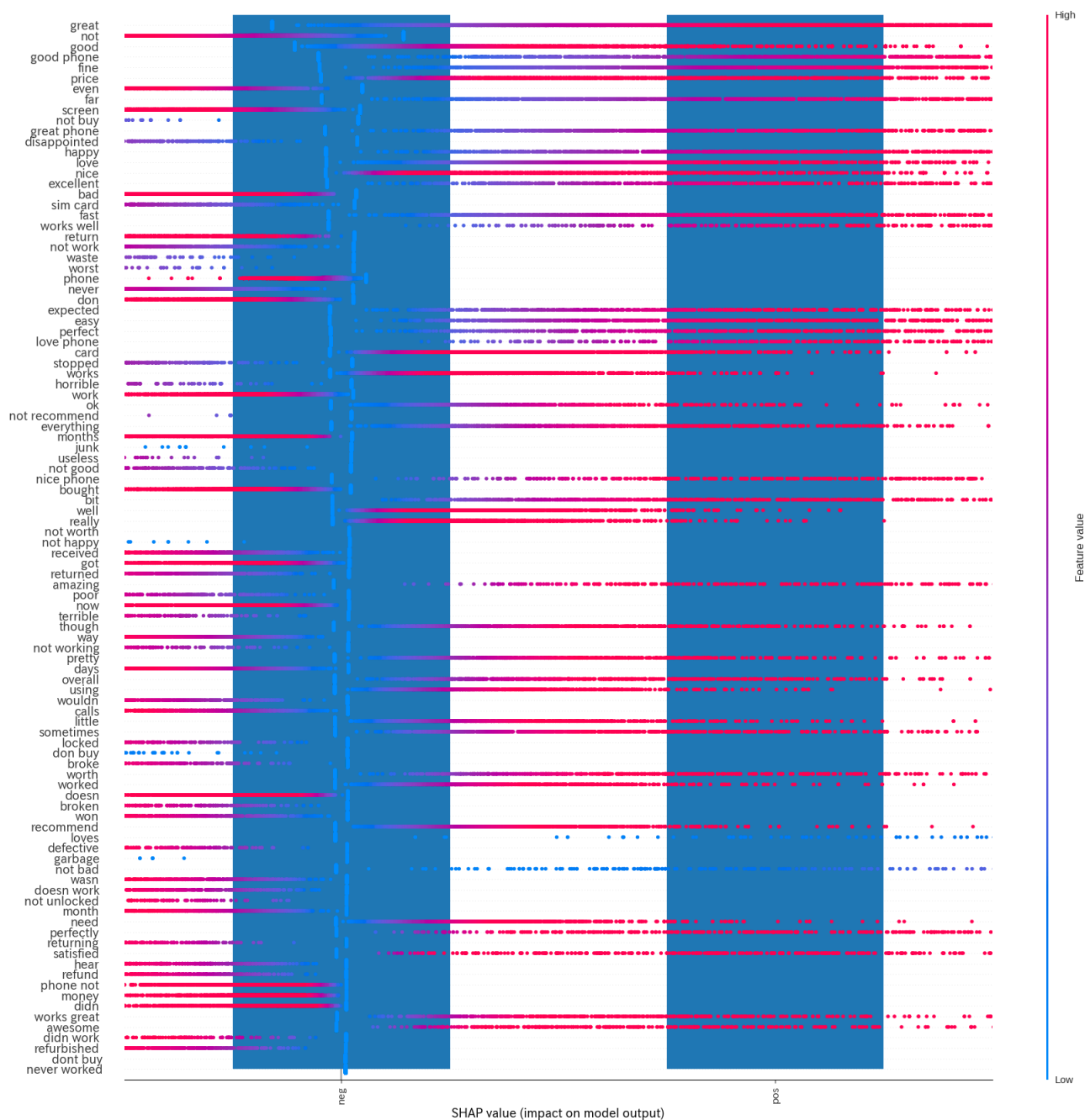


図 7 2gram での SHAP 値の高い語のリスト.

表 5 ロジスティック回帰による手書きレビューの感情予測（ネガティブの例）

レビューの内容	予測した感情	ネガティブ確率	ポジティブ確率
"product is not good"	NEG	0.996505	0.003495
"I will never buy this again"	NEG	0.999957	0.000043
"not good"	NEG	0.989669	0.010331