



Bag of Words Meets Bags of Popcorn

Use Google's Word2Vec for movie reviews

Getting Started · 577 Teams · 11 years ago

<https://www.kaggle.com/competitions/word2vec-nlp-tutorial>

古典から最新へ：AIエージェントとローカルGPUで切り拓く最高精度 AUC 0.98393 の軌跡

Bag of Words Meets Bags of Popcorn コンペティション
実験レポートと成功のプレイブック

庭師

- Gemini, Notebook LM, ちよっとだけChatGPT
- Kaggle環境で実行, 自PC(Core i7-11700K+RTX3060(12GB))



この資料を作成した方法と内容

・ 作成方法

- ・ GeminiとChatGPTで、コンペ概要、EDAを実施。
 - ・ どちらも65MB級データは読み込みめた。
- ・ 比べた結果、Geminiの方が解析能力が良さそう。
 - ・ Geminiを選択。
- ・ pptxに経緯を記録
- ・ Geminiに最高スコアを狙うコーディングさせる。
- ・ 最初はKaggle Notebook環境使っていたが、T4 x2環境よりも、自PC(Core i7-11700K+RTX3060)の方が高速だとわかり、自PCメインに切り替え。
- ・ Geminiに経緯記録したpptxもuploadして、レポート作成。
- ・ そのレポートをNotebook LMに渡してスライド作成

・ レポートの内容

- ・ コンペ紹介、概要
- ・ EDA、高スコアを出すコツ
- ・ コーディング説明、推移
- ・ まとめ

Claude Codeの首位陥落。CodexがシェアNo.1へ。

- Claude Codeの首位陥落。CodexがシェアNo.1へ。～データで見る2026年3月のAI Codingの動向まとめ～ 2026/04/04

• <https://qiita.com/Kotauchisunsun/items/ab78bb338500b4c71103>

• 2026年2月中旬、国防総省がアンソロピックのAIソフトの軍事利用において、最新の提案を拒否したことが報じられました。

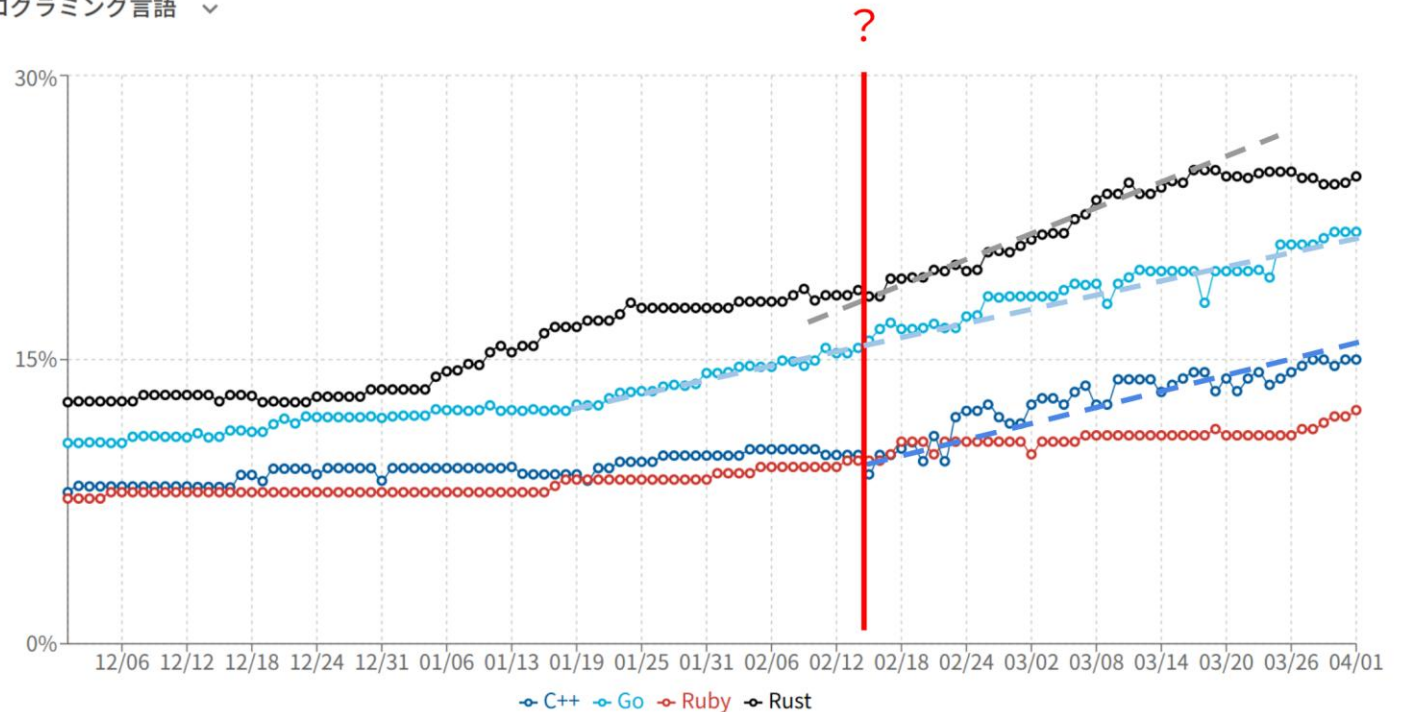
**Codexは、ChatGPT plusで使える。
Claude Codeより安そう！**

でも今回は Gemini で実行

AI Coding Agentの利用率の推移 (2025年12月1日-2026年4月1日) ⇄

リポジトリのAI Coding Agentの利用率の推移を表しています。

プログラミング言語 ▾





コンペを説明して

Submission Limits

You may submit a maximum of 5 entries per day.

でもlate submissionなので100エントリー/日出来るようだ

- このコンペティション「**Bag of Words Meets Bags of Popcorn**」は、Kaggleが自然言語処理(NLP)の初心者向けに提供している「**Getting Started**」コンペの一つです。
- 映画のレビュー文を読んで、それが「**肯定的(ポジティブ)**」か「**否定的(ネガティブ)**」かを判定する**感情分析(Sentiment Analysis)**をテーマにしています。
- 以下に、コンペの概要、データ、学ぶべき主要技術をまとめました。

1. コンペの概要

- **目的:** IMDB(映画データベース)のレビューテキストを解析し、そのレビューが**ポジティブ(1)**か**ネガティブ(0)**かを予測する。
- **形式:** 2値分類問題。
- **評価指標:** **AUC (Area Under the ROC Curve)**。予測の正確さを測る指標です。

2. データセットの内容

データは非常にシンプルで扱いやすい構成になっています。

- **学習データ (labeledTrainData):** 25,000件のレビュー。ID、感情ラベル(0 or 1)、レビュー本文が含まれます。
- **テストデータ (testData):** 25,000件のレビュー。感情ラベルは伏せられており、これを予測して提出します。
- **ラベルなし追加データ (unlabeledTrainData):** 50,000件のレビュー。Word2Vecなどの「教師なし学習」に利用するために提供されています。

3. このコンペで学べる主要な手法

このコンペはタイトルにある通り、伝統的な手法(Bag of Words)と、当時革新的だった手法(Word2Vec)の両方を比較・学習できるよう設計されています。

① Bag of Words (BoW)

文章を「単語の袋」として扱い、どの単語が何回出てきたかをカウントしてベクトル化する手法です。

- **特徴:** シンプルで強力ですが、単語の順序や意味(文脈)は無視されます。
- **よく使われるモデル:** Random Forest, Logistic Regressionなど。

② Word2Vec (Googleの手法)

単語を多次元のベクトルで表現し、単語同士の意味的な近さを計算できるようにする手法です。

- **特徴:** 「王様 - 男性 + 女性 = 女王」のような計算ができることで有名です。文脈をある程度考慮した特徴量を抽出できます。
- **このコンペでの役割:** ラベルのない大量のデータを使って単語ベクトルを学習し、それを分類モデルに活用する方法を学びます。

4. 進め方のステップ(チュートリアル形式)

Kaggle公式が丁寧なチュートリアルを用意しており、以下の順にステップアップできます。

1. **Part 1:** 基本的なテキストクリーニング(HTMLタグ削除、ストップワード除去)と Bag of Words を使った分類。
2. **Part 2:** Word2Vec を使って独自の単語ベクトルを学習させる。
3. **Part 3:** 学習したベクトルを使って、文章全体のベクトルを作り(平均化など)、分類器にかける。





5. なぜこのコンペがおすすめか?

- **NLPの基礎が詰まっている:** テキストの正規化(BeautifulSoup やNLTKの使用)から、最新の深層学習への入り口まで一通り体験できます。
- **データが綺麗:** 映画レビューという、感情がはっきり出やすいテキストなので、初心者でも比較的高い精度が出しやすく、達成感があります。

自然言語処理をこれから始めたい人にとって、登竜門として最適なコンペティションです。まずは Python の `scikit-learn` で BoW を試すところから始めるのが王道です

データサイズ

- 一番大きいので65MB。
- このサイズだと、ChatGPTでも、Geminiでもデータをuploadできた。

名前	サイズ
 sampleSubmission.csv	277 KB
 labeledTrainData.tsv	32,770 KB
 testData.tsv	31,958 KB
 unlabeledTrainData.tsv	65,705 KB

データの一例: labeledTrainData.tsv

```

1 id sentiment review
2 "5814_8" 1 "With all this stuff going down at the moment with MJ i've started listening to h
3 "2381_9" 1 "\"The Classic War of the Worlds\" by Timothy Hines is a very entertaining film t
4 "7759_3" 0 "The film starts with a manager (Nicholas Bell) giving welcome investors (Robert
5 "3630_4" 0 "It must be assumed that those who praised this film (\\"the greatest filmed opera
6 "9495_8" 1 "Superbly trashy and wondrously unpretentious 80's exploitation, hooray! The pre-
7 "8196_8" 1 "I dont know why people think this is such a bad movie. Its got a pretty good plo
8 "7166_2" 0 "This movie could have been very good, but comes up way short. Cheesy special eff
9 "10633_1" 0 "I watched this video at a friend's house. I'm glad I did not waste money buying
10 "319_1" 0 "A friend of mine bought this film for £1, and even then it was grossly overpriced. D
11 "8713_10" 1 "<br /><br />This movie is full of references. Like \\"Mad Max II\\", \\"The wild on
12 "2486_3" 0 "What happens when an army of wetbacks, towelheads, and Godless Eastern European
13 "6811_10" 1 "Although I generally do not like remakes believing that remakes are waste of tim
14 "11744_9" 1 "\\"Mr. Harvey Lights a Candle\" is anchored by a brilliant performance by Timothy
15 "7369_1" 0 "I had a feeling that after \\"Submerged\\", this one wouldn't be any better... I w
16 "12081_1" 0 "note to George Litman, and others: the Mystery Science Theater 3000 riff is \\"I
17 "3561_4" 0 "Stephen King adaptation (scripted by King himself) in which a young family, newc
18 "4489_1" 0 "\\"The Matrix\" was an exciting summer blockbuster that was visually fantastic but
19 "3951_2" 0 "Ulli Lommel's 1980 film 'The Boogey Man' is no classic, but it's an above averag
20 "3304_10" 1 "This movie is one among the very few Indian movies, that would never fade away w
21 "9352_10" 1 "Most people, especially young people, may not understand this film. It looks lik
22 "3374_7" 1 "\\"Soylent Green\" is one of the best and most disturbing science fiction movies
23 "10782_7" 1 "Michael Stearns plays Mike, a sexually frustrated individual with an interesting
24 "5414_10" 1 "This happy-go-luck 1939 military swashbuckler, based rather loosely on Rudyard K
25 "10492_1" 0 "I would love to have that two hours of my life back. It seemed to be several cli
26 "3350_3" 0 "The script for this movie was probably found in a hair-ball recently coughed up
27 "6581_7" 1 "Looking for Quo Vadis at my local video store, I found this 1985 version that lo
28 "2203_3" 0 "Note to all mad scientists everywhere: if you're going to turn your son into a g

```

```

1 id sentiment review↓
2 "5814_8" 1 "With all this stuff going down at the moment with MJ i'
3 ve started listening to his music, watching the odd documentary here and there,
4 watched The Wiz and watched Moonwalker again. Maybe i just want to get a certain
5 insight into this guy who i thought was really cool in the eighties just to may
6 be make up my mind whether he is guilty or innocent. Moonwalker is part biograph
7 y, part feature film which i remember going to see at the cinema when it was ori
8 ginally released. Some of it has subtle messages about MJ's feeling towards the
9 press and also the obvious message of drugs are bad m'kay.<br /><br />Visually i
10 mpressive but of course this is all about Michael Jackson so unless you remotely
11 like MJ in anyway then you are going to hate this and find it boring. Some may
12 call MJ an egotist for consenting to the making of this movie BUT MJ and most of
13 his fans would say that he made it for the fans which if true is really nice of
14 him.<br /><br />The actual feature film bit when it finally starts is only on f
15 or 20 minutes or so excluding the Smooth Criminal sequence and Joe Pesci is conv
16 incing as a psychopathic all powerful drug lord. Why he wants MJ dead so bad is
17 beyond me. Because MJ overheard his plans? Nah, Joe Pesci's character ranted tha
18 t he wanted people to know it is he who is supplying drugs etc so i dunno, maybe
19 he just hates MJ's music.<br /><br />Lots of cool things in this like MJ turnin
20 g into a car and a robot and the whole Speed Demon sequence. Also, the director
21 must have had the patience of a saint when it came to filming the kiddy Bad sequ
22 ence as usually directors hate working with one kid let alone a whole bunch of t
23 hem performing a complex dance scene.<br /><br />Bottom line, this movie is for
24 people who like MJ on one level or another (which i think is most people). If no
25 t, then stay away. It does try and give off a wholesome message and ironically M
26 J's bestest buddy in this movie is a girl! Michael Jackson is truly one of the m
27 ost talented people ever to grace this planet but is he guilty? Well, with all t
28 he attention i've gave this subject...hmmm well i don't know because people can
29 be different behind closed doors, i know this for a fact. He is either an extre
30 mely nice but stupid guy or one of the most sickest liars. I hope he is not the
31 latter.↓
32 "2381_9" 1 "\"The Classic War of the Worlds\" by Timothy Hines is a
33 very entertaining film that obviously goes to great effort and lengths to faith
34 fully recreate H. G. Wells' classic book. Mr. Hines succeeds in doing so. I, and
35 those who watched his film with me, appreciated the fact that it was not the st
36 andard, predictable Hollywood fare that comes out every year, e.g. the Spielberg
37 version with Tom Cruise that had only the slightest resemblance to the book. Ob
38 viously, everyone looks for different things in a movie. Those who envision them
39 selves as amateur ¥critics¥ look only to criticize everything they can. Others
40 rate a movie on more important bases, like being entertained, which is why most
41 people never agree with the ¥critics¥. We enjoyed the effort Mr. Hines put int
42 o being faithful to H.G. Wells' classic novel, and we found it to be very entert
43 aining. This made it easy to overlook what the ¥critics¥ perceive to be its sh
44 ortcomings.↓

```

値	意味 (Sentiment)	内容の例
1	Positive (肯定的)	「面白い」「感動した」「俳優の演技が素晴らしい」「また見たい」など
0	Negative (否定的)	「つまらない」「時間の無駄」「脚本がひどい」「がっかりした」など

sentiment 1, 0 の意味は？

- このコンペにおける sentiment の 1 と 0 は、映画レビューの感情が「肯定的」か「否定的」かを示すラベルです。
- **なぜこれが重要なのか？**
- このコンペの目的は、学習データ (labeledTrainData.tsv) を使って、「1 か 0 か」を予測するモデルを作ることです。
- **1 (Positive) の予測確率:** モデルが「このレビューは 80% の確率で 1 だ」と判断すれば、そのレビューはポジティブ寄りであると解釈されます。
- **評価指標 (AUC):** このコンペでは、単に「1か0か」を当てるだけでなく、**「どれくらい自信を持って1(または0)と言えるか」** という確率の正確さがスコア (AUC: Area Under the ROC Curve) に直結します。
- **データの傾向 (アップロードされたファイルより)**
- 提供された labeledTrainData.tsv を見ると、以下のような特徴があります。
- **IDの法則:** id が "5814_8" のようになっていますが、アンダースコアの後の数字 (この場合は 8) は、元の IMDB での 10 点満点の評価スコアです。
 - **スコア 7 以上** → sentiment: 1 (Positive)
 - **スコア 4 以下** → sentiment: 0 (Negative)
 - ※ 5 点 ~ 6 点の中間的なレビューは、あえてデータから除外されており、感情がはっきり分かれるようになっています。
- **最高スコアを出すためには、単に単語を拾うだけでなく、これら「極端な感情表現」をモデルにいかに学習させるかがカギになります！**

古典から最新へ：AIエージェントとローカルGPUで切り拓く最高精度 AUC 0.98393 の軌跡

Bag of Words Meets Bags of Popcorn コンペティション
実験レポートと成功のプレイブック

AUC 0.98393

感情分析の登竜門：IMDb映画レビュー分類タスクの全貌

The Task (タスク)



映画のレビュー文の感情を判定する2値分類タスク。

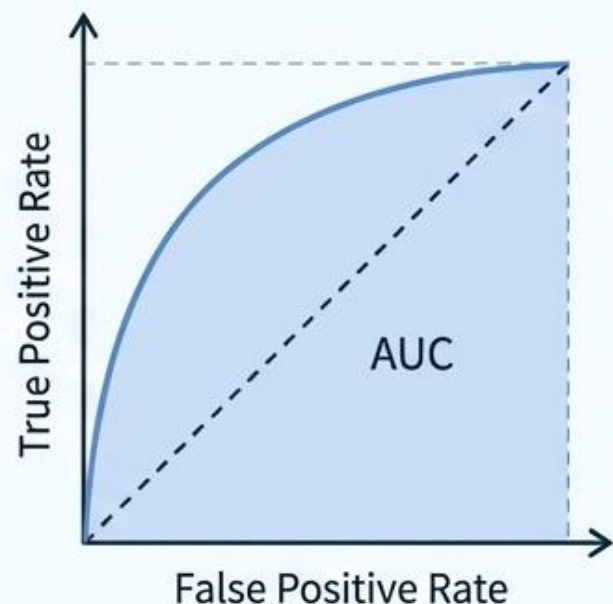
The Dataset (データセット)

```
1 id sentiment review
2 "5814_8" 1 "With all this stuff going down at the moment wi
3 "2381_8" 1 "\"The Classic War of the Worlds\" by Timothy Hir
4 "7759_3" 0 "The film starts with a manager (Nicholas Bell) g
5 "3630_4" 0 "It must be assumed that those who praised this f
6 "9495_5" 1 "Superbly trashy and wondrously unpretentious 88'
7 "8156_5" 1 "I dont know why people think this is such a bad
8 "7166_2" 0 "This wovid could have been very good, but comes
9 "19633_1" 0 "I watched this video at a friend's house. I'm gi
10 "219_1" 0 "A friend of mine bought this film for £1, and even t
11
12
13
14
15
16 "3561_1" 0 "Stephen King adaptation (scripted by King himsel
17 "4489_1" 1 ""The Matrix" was an exciting summer blockbuster
18 "3951_3" 1 "Ulli Lemmel's 1980 film "The Boogey Man" is no c
19 "3374_18" 1 "This novie is one among the very few Indian wovid
20 "3374_7" 1 "Michaseet Green's mlke, of perposo, and nest di
21 "5414_18" 1 "This happy-go-luck 1939 military swashbuckler, ba
22 "18492_0" 0 "I would love to have that two hours of my life ba
23 "3351_3" 1 "Looking for Quo Vadis at my local video store, I
24 "2283_3" 0 "Note to all mad scientists everyuhere: if you're
```

50% 50%

学習用: 25,000件 / テスト用:
25,000件 / 追加未ラベル: 50,000件
(ポジティブ/ネガティブ完全均衡)

The Metric (評価指標)

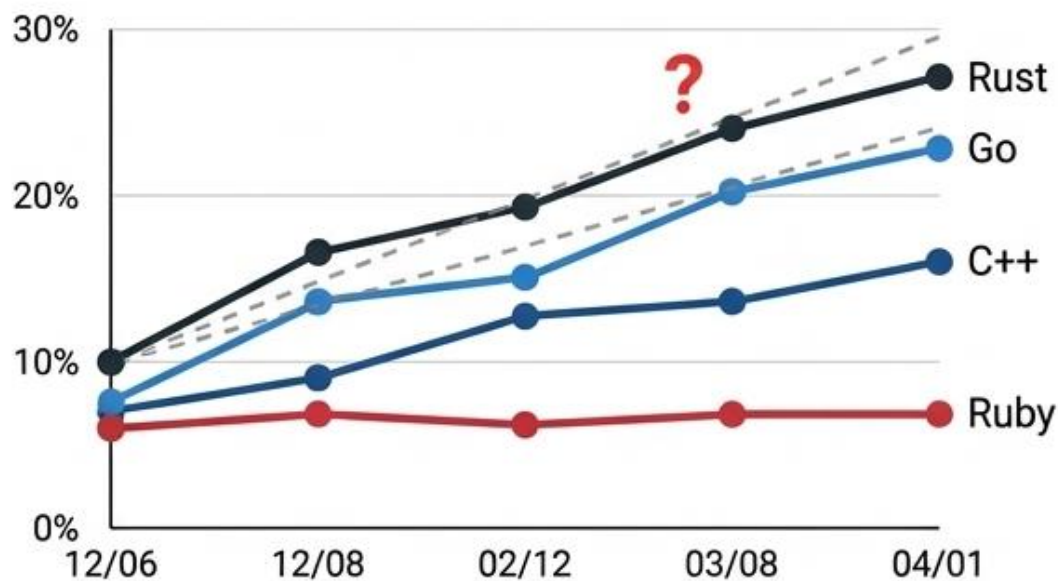


AUC (Area Under the ROC Curve)。0か1かの正解だけでなく、「どれくらい自信を持って予測できたか」の確率の正確さが問われる。

高速イテレーションを生み出した「AI相棒」と「ローカル環境」の相乗効果

AI Coding Agents

メイン分析とコーディングに「Gemini」を採用。
65MB級データの高速読み込みと高度な解析能力が決め手。



AI Coding Agentの利用率の推移
(2025年12月1日-2026年4月1日)

Hardware Face-off: Cloud vs Local

Kaggle T4 x2 (Cloud)

約8時間 (途中で中止)

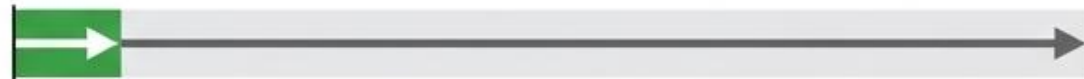


0

約8時間 (途中で中止)

自作PC (Core i7 + RTX 3060 12GB)

25分~1時間強



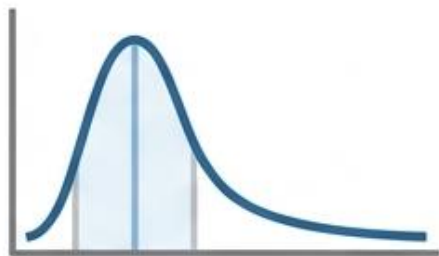
0

25分~1時間強

Insight: コンペの勝敗は「試行回数」で決まる。
VRAM 12GBのローカルPCへの回帰が、最大の武器となった。

探索的データ分析 (EDA) で見えた3つの「勝ち筋」と戦略的急所

レビュー長の分布



平均230単語だが、1,000語超えの長文も多数存在。

Transformerの
MAX_LEN 設定 (256または384) が極めて重要。

HTMLタグのノイズ



レビュー内に `
` などのWeb特有のノイズが散見される。

BeautifulSoup等を用いた事前のテキストクリーニングが必須。

否定表現の重み



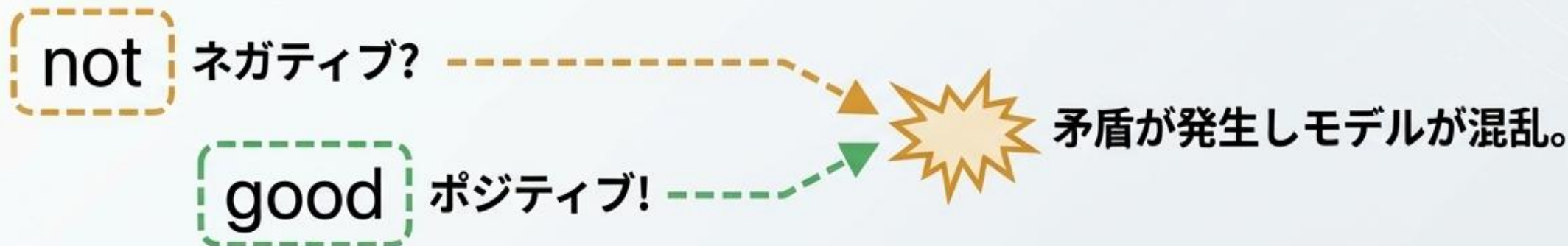
「not」「never」などの極端な感情・否定表現が結果を大きく左右する。

単純な単語カウントではなく、文脈・表現の検出能力がスコアの鍵を握る。

「単語」から「文脈」へ：否定表現がスコアの分水嶺となる理由

This movie is not good

Unigram / Bag of Words の限界



N-gram / Transformer の文脈理解力



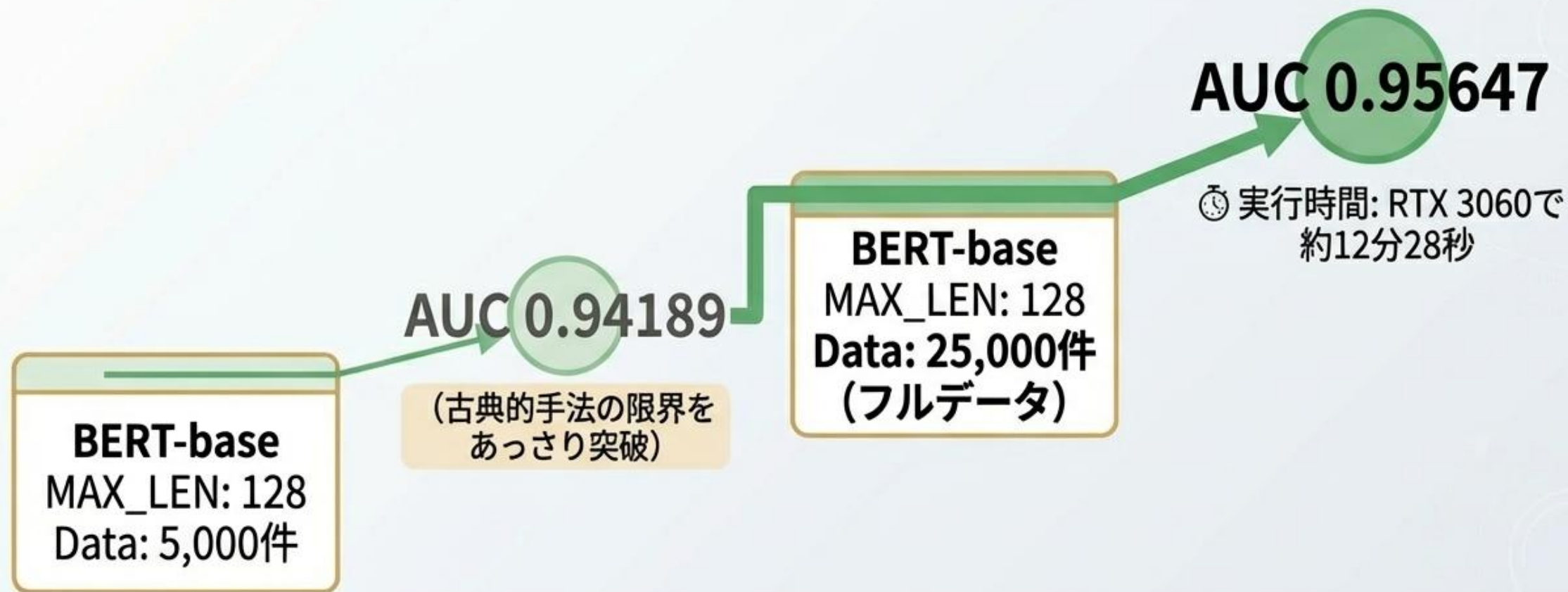
Insight: 「goodよりnot goodの方が情報量が多い」。
これがBoW (LB 0.85) と文脈理解モデル (LB 0.90超) の壁。

どのモデルで戦うか：NLPアーキテクチャの進化と比較マトリクス

アーキテクチャ	難易度	期待AUC	特性と評価
Word2Vec + Random Forest	低	~0.84	チュートリアル標準。文脈理解は限定的。
TF-IDF + Naive Bayes	低	~0.90	シンプルだがN-gramとの組み合わせで異常に強い。
Doc2Vec + MLP	中	~0.93	文書全体のベクトル化。
Transformer (BERT / DeBERTa)	高	0.96~0.99+	現代の最高峰。文脈の双方向理解が可能だが計算資源が必須。

本コンペの最適解

Stage 1 : ベースライン突破と「データ量の暴力」



Insight: AIにとってデータ量は「経験値」。フルデータ学習により、未知のデータへの汎化性能が劇的に向上し、Topクラスの土台が完成。

Stage 2: オチまで読み切る「コンテキストの拡張」



New Score:
Validation Green
AUC 0.97483
Top 1~5%圏内へ突入
🕒 実行時間: 約25分15秒

Stage 3 : "人間超え"の感情理解。最高精度への到達点

最適解

Model: RoBERTa-base

MAX_LEN: 256

Learning Rate: 1e-5



Zenith

AUC 0.98393

Time: RTX 3060 / 25m58s

**Insight: 複雑な皮肉やニュアンスを完全に捉え、
精度・学習時間・VRAM制約の最も完璧なバランスを達成した瞬間。**

Stage 4 : 失敗から学ぶ「最新・最大」が招いたオーバースペックの罠

Failure 1: DeBERTa-v3-base の崩壊

Result:

Loss=NaN エラー発生 (27分で破綻)

Cause:

相対位置エンコーディングなどの特殊構造が、現在の環境設定下で数値計算の不安定化を引き起こした。

Failure 2: RoBERTa-large の空回り

Result:

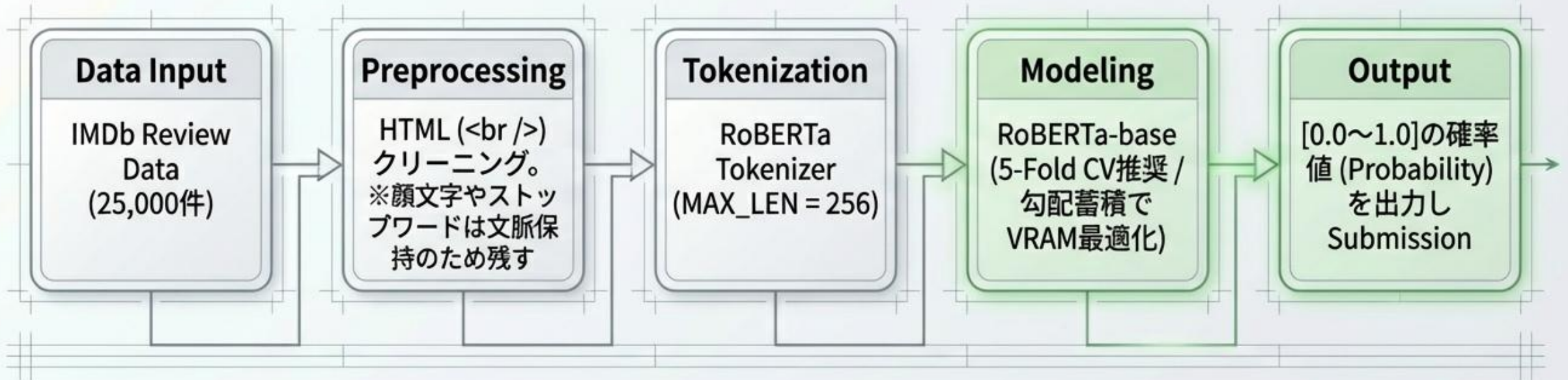
AUC 0.95064 (大幅なスコアダウン) / 実行時間 111分

Cause:

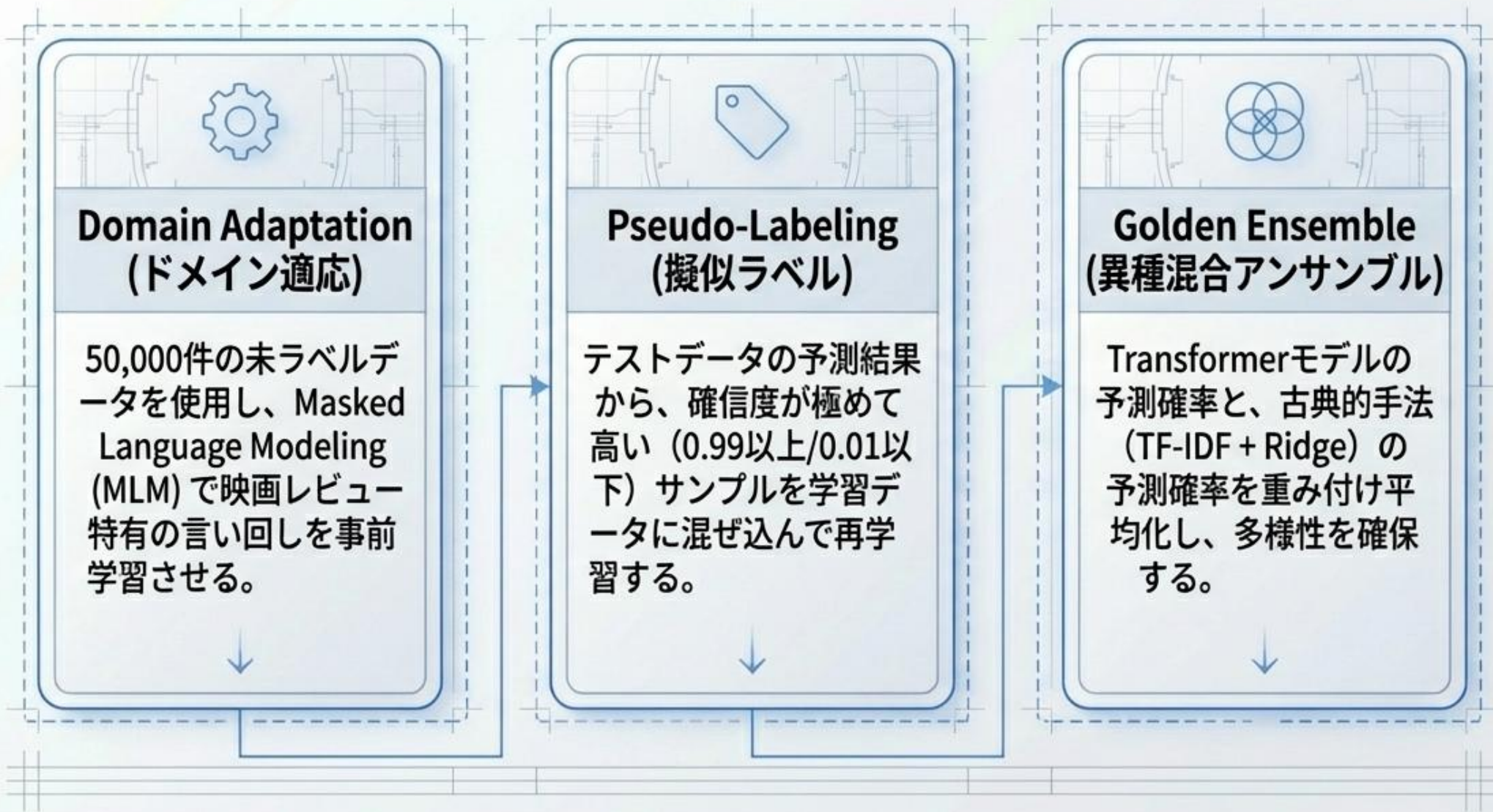
3.5億パラメータという巨大モデルに対し、2.5万件のデータでは学習不足・過学習のミスマッチが発生。

Takeaway: 「大きければ良い」わけではない。ハードウェア規模とデータセットに合わせた「最適なバランス (Baseサイズ)」を見極めることが勝利の鍵。

最高スコアを生み出した最終アーキテクチャの青写真



限界突破の次の一手：AUC 0.99超えを狙う高度な戦略



結論とキーテイクアウェイ：実践から得られた3つの教訓

1

文脈の長さは正義 (Context is King)

NLPタスクにおいて、テキストの末尾（結論）を切り捨てない MAX_LEN の設定は、モデル選びと同等以上に重要である。

2

バランスの勝利 (The Right Fit)

最新最大のモデルが勝つとは限らない。精度・学習時間・ハードウェア制約のバランスが取れたモデル (RoBERTa-base) こそが最適解となる。

3

ローカル×AIの爆発力 (The Local Advantage)

クラウドに縛られず手元の RTX 3060 環境を活かし、AIエージェントと協働して「試行回数」を回しきったことが Top 水準到達の最大の原動力。