

# 1on1 ミーティング時のメモを用いたワークメンタリティの好不調の推定

## Estimating Work Mentality from One-on-one Meeting Memo

宇野 渉\*<sup>1</sup>  
Wataru Uno

仲間 大輔\*<sup>1</sup>  
Daisuke Nakama

\*<sup>1</sup> 株式会社リクルートマネジメントソリューションズ  
Recruit Management Solutions Co., Ltd.

It is an important issue for HR managers to assess the work mentality of employees to provide appropriate support, for the sake of better human capital management. While employee surveys about their own work mentality is commonly used, it is often difficult to conduct such surveys frequently due to the burden on the employees. Thus, it will be practically useful if mentality can be estimated without conducting employee surveys. In this study, we developed models to estimate the work mentality of employees from memos from “one-on-one” meetings held between a manager and a member, which have been popular in recent years. With 3186 memos from actual one-on-one meetings and results from mentality surveys, we compared three machine learning models, including a model with Doc2Vec, BERT-sentiment, and fine-tuned BERT. The highest performance model was fine-tuned BERT with F1 0.69 and Accuracy 0.75. Furthermore, we compared the classification performance between fine-tuned BERT and human estimations from the same memos. Results indicated that fine-tuned BERT outperformed non-experts and was comparable with an expert.

### 1. はじめに

企業経営における人的資本の重要性が増すと同時に、人材不足が深刻化する中、職場における感情の重要性に注目が集まっている[Cropanzano 2017]。社員一人一人のメンタリティを把握し、適切にサポートすることは重要な人事課題となっているが[Weiss 1996]、これに対する施策として、社員に今の自身の状態について数問から数十問の設問に回答してもらうアンケートの実施が一般的である。これを定期的の実施することにより、社員一人一人のメンタリティをモニタリングすることが期待できる。その一方で、回答する社員の負荷が高いため高頻度の実施が難しいといった問題がある。そのため、アンケートに依らずに従業員のメンタリティを推定することができれば、補完的に用いることが出来るため、実務上の価値が高い。

ここで、最近では上司と部下が 1 対 1 で定期的に対話する 1on1 ミーティングが多くの企業で導入されていることに注目する[リクルートマネジメントソリューションズ 2022]。1on1 ミーティングでは普段一緒に仕事をしている上司が、直接部下であるメンバーから業務の話だけでなく、将来のキャリアの話やプライベートの話も聴く機会である。上司は複数のメンバーと 1on1 ミーティングを実施するため、今回の 1on1 ミーティングで何を会話をしたかといったメモ(以下: 1on1 メモ)を記録し、そのメモを人事に共有するといったことが一般的に行われる。これらのメモには、上司が 1on1 ミーティング中の対話から得られたメンバーのメンタリティに関する情報が含まれるため、メンバーのメンタリティを推定するのに有用であると考えられる。

そこで本研究では、1on1 メモを用いてメンバーのメンタリティを推定する機械学習モデルを複数作成し、どの程度の精度で推定できるのかを検証する(2 章、3 章)。その上で、モデルの精度が人間による判定と比べてどの程度のものであるのかを調べるため、人による推定実験を実施した上で比較検証を行った(4 章)。

### 2. 提案手法

#### 2.1 1on1 メモデータ

本研究では 1on1 の支援を目的として提供されている Web サービス上で、実際に上司が記録した 1on1 メモを用いる。上司は、メンバーとの 1on1 ミーティング中もしくは終了後に自分用のメモと人事への共有用のメモを記録する。本研究では、このうち共有用のメモを推定に用いる。

また、同サービス内には、社員が仕事に向かう心理状態(以下:ワークメンタリティ)についての従業員アンケート結果も格納されている。このアンケートによるワークメンタリティの回答結果を、1on1 メモから推定する対象の目的変数とする。1on1 メモとワークメンタリティの紐づけについては、1on1 メモに記載されている面談日に最も近いアンケート結果のワークメンタリティを紐づける。ただし、面談日とアンケート回答日の間の期間は最大 30 日とし、それを超えるものは使用しないことにした。アンケート回答は 5 段階スケールであったため、特に職場でのサポートが必要と考えられる下位 2 段階を不調、その他を好調と定義し、2 値の分類タスクとして扱った。

表 1 に不調・好調のメンバーそれぞれについて上司が記録した 1on1 メモの例を示す。表 1 中の文章は実際のメモではなく、実際のメモを参考に著者が作成したものである。

表 1: 1on1 メモの例

メンタリティ	1on1 メモ文章
好調	仕事量に関してはまだ余裕があり、全般的に仕事に楽しく取り組んでいる。余裕のあるうちにもっと大きな仕事や勉強がしたいとのこと。
不調	自身の業務をこなすだけで精一杯になっており、グループの人員も不足しており業務がまったく回っていない。新しいことにチャレンジできない環境に不満がある。

## 2.2 推定モデル

本研究では、1on1 メモからワークメンタリティの好不調を推定するために、3つのモデルを使用した。

### (1) Doc2Vecによりベクトル化した Xgboost

1つ目は Doc2Vec[Le 2014]を使用したモデル(以下、D2V-xgb)である。このモデルでは、各 1on1 メモを 300 次元のベクトルに変換したものを特徴量として Xgboost[Chen 2016]によるモデリングを行った。特徴的なワードのみで推定できる場合は以降 2つのモデルと同程度の精度が期待できる。

### (2) 感情分類 BERT モデル

2つ目は BERT[Devlin 2018]の事前学習モデルをベースに作成された感情分類モデル(以下、BERT-sentiment)である。こちらは、Hugging Face に公開されている jarvisx17/japanese-sentiment-analysis を使用した。このモデルは、上場企業の有価証券報告書の各文にポジティブ/ネガティブフラグが付与されたデータセットをもとに作成されたモデルである。本研究では、このモデルでポジティブと判定されたものを好調、ネガティブと判定されたものを不調と定義した。このモデルは 1on1 メモをベースにチューニングされたものではないが、ワークメンタリティ不調者について書かれたメモについても共通するネガティブな表現がある場合には、感情分類モデルでも一定の精度で推定できると期待できる。

### (3) 1on1 メモによりファインチューニングした BERT モデル

3つ目は BERT の事前学習モデルを、1on1 メモとそれらに紐づけられたワークメンタリティラベルのデータによりファインチューニングしたモデル(以下、BERT-finetuned)である。日本語の事前学習済みモデルには Hugging Face に公開されている cl-tohoku/bert-base-japanese-v2 を使用した。このモデルの構成は 12 層の Layer、768 次元の Hidden States、12 個の Attention Head から構成される。BERT-sentiment は 1on1 メモでないデータでチューニングされているのに対し、この BERT-finetuned は実際の 1on1 メモによる学習をするため、1on1 メモ特有の表現などの特徴がある場合には、BERT-sentiment よりも高精度が期待できる。

## 3. 推定モデルの精度比較

### 3.1 データセット

本研究では 2.1 で上述した、1on1 の支援を目的として提供されている Web サービス上で、実際に上司が記録した 1on1 メモを用いる。データ数は 3186 レコードで、854 人によって記録されたものである。目的変数となるワークメンタリティについては、3186 レコードのうち、1345 レコード(約 42.2%)が不調者のメンバーについて書かれたものである。

本検証では、全データの 2 割をテストデータ、残りの 8 割を訓練データになるように分割した。なお、分割ではテストデータと訓練データで目的変数の割合が同じになるように層化分割している。

### 3.2 3モデルの精度比較

#### (1) 精度比較結果

精度比較結果を表 2 に示す。分類精度の評価尺度として Accuracy、Precision、Recall、F1-Score を使用している。F1-

Score で比較すると、BERT-finetuned が最も高精度で、その次に BERT-sentiment という結果になった。

表 2: 3モデルの精度比較結果

Model	F1-Score	Accuracy	Precision	Recall
D2V-xgb	0.35	0.56	0.29	0.46
BERT-sentiment	0.54	0.52	0.45	0.67
BERT-finetuned	0.69	0.75	0.65	0.73

#### (2) 考察

まず、D2V-xgb と BERT の 2 モデルで大きく精度に差異が生じたことについて考察する。表 3 に示す 2 つの 1on1 メモは、D2V-xgb と BERT の 2 モデルにおいて、両方が正解したものと、BERT の 2 モデルでのみ正解したものを参考に、著者が作成したものである。なお、作成後に各モデルで予測し、推定結果が変わらないことを確認している。以下、著者が作成した 1on1 メモについては同様の確認をしている。

D2V-xgb が正解した 1on1 メモでは全般的に、「ストレス」「忙しい」など明らかにメンバーが不調であることが推測できる特徴語があった。その一方で、D2V-xgb が不正解で BERT の 2 モデルが正解した 1on1 メモでは「元気」「成果」「満足」など好調なメンバーの 1on1 メモでも頻出する単語が含まれており、単語だけでは好不調を推定できない表現になっていた。そのため、文脈を考慮しない D2V-xgb では BERT に比べて精度が低かったと考えられる。

表 3: 1on1 メモの例と D2V-xgb と BERT の推定結果

1on1 メモ文章	D2V-xgb	BERT
仕事があまくいかないことにストレスを感じている様子です。仕事量が多く、残業も多いため、とても忙しいとのこと。[不調]	不調 (正解)	不調 (正解)
これまでは元気そうで成果も出していました、本人は満足できていないようです。引き続きフォローしていこうと思います。[不調]	好調 (不正解)	不調 (正解)

次に、BERT-sentiment よりも BERT-finetuned の方が高精度となった要因について考察する。表 4 に BERT-sentiment で不正解で、BERT-finetuned で正解だった 1on1 メモを参考に著者が作成したものを例として示す。この例では、「仕事量が多い」や「モチベーションが上がらない」などの表現から不調と推定される。しかし、BERT-sentiment では好調と判定されている。これはこういった表現が 1on1 メモ特有の表現であると考えられるため、別のデータセットで作成された BERT-sentiment では不調と推定できなかったと考えられる。

表 4: 1on1 メモの例と各 BERT モデルの推定結果

1on1 メモ文章	BERT-sentiment	BERT-finetuned
異動前は仕事量が多かったため業務で手一杯だったが、異動後に仕事量が減っておりモチベーションがなかなか上がらない様子。[不調]	好調 (不正解)	不調 (正解)

### 3.3 まとめ

1on1 メモによりファインチューニングしたモデル (BERT-finetuned) が最も高い精度を示した。これは、実務上の応用可能性を示唆しているが、今回示された精度 (F1 で 0.69) が、実際にどのような水準となっているのかをベンチマークを用いて検証することが必要と考えられる。1on1 メモでは必ずしもメンバーのワークメンタリティを示唆する記述があるわけではないため、人が推定したとしてもすべての 1on1 メモからワークメンタリティをすべて正確に予測できるわけではない。そこで次章において、人による推定との比較実験を行う。

## 4. 人による推定精度との比較

最も高精度だった BERT-finetuned について、人による推定精度との比較を行った。

### 4.1 実験環境

人による推定では合計 9 人の被験者 (S1~S8, E1) の協力を得た。うち 1 名 (E1) は特に、本研究で使用しているデータが記録されている Web サービスのカスタマーサクセスのリーダーであり、1on1 について精通しているエキスパートである。

各被験者には、各自で 1on1 メモを読んでもらい、「対象の 1on1 メモが書かれた面談のときに面談対象のメンバーのワークメンタリティが不調だったかどうか」を推定してもらった。データについては 3 章で使用したテストデータから 200 件を層化サンプリングしたものを使用した。そのため、BERT-finetuned の精度についても表 2 に示したものと異なるため、併せて表 3 に記載する。

精度比較では、各被験者の精度だけでなく、非専門家 (S1~S8) の多数決によるアンサンブルとの精度についても比較した。なお、アンサンブルでは、各自に重みづけなどはせず、半数以上の推定を採用している。

### 4.2 精度比較結果

各被験者 (S1~S8, E1) とアンサンブル (Ensemble)、BERT-finetuned の精度を表 5 に示す (F1-Score で降順にソートされている)。表 5 が示す通り、BERT-finetuned はほとんどの被験者よりも高精度であり、エキスパートである E1 と同程度の精度で推定できることが明らかになった。

表 5: 人による推定精度との比較結果

Subject/Model	F1-Score	Accuracy	Precision	Recall
<b>E1 (Expert)</b>	0.71	0.74	0.67	0.74
<b>BERT-finetuned</b>	0.70	0.76	0.72	0.68
S7	0.67	0.63	0.53	0.89
<b>Ensemble</b>	0.64	0.69	0.63	0.64
S6	0.64	0.70	0.64	0.63
S3	0.60	0.69	0.66	0.55
S2	0.58	0.68	0.63	0.54
S1	0.56	0.67	0.63	0.50
S8	0.55	0.69	0.69	0.45
S4	0.55	0.60	0.52	0.57
S5	0.52	0.61	0.54	0.50

上述の精度比較結果について、BERT-finetuned、エキスパートである E1、S1~8 のその他の被験者 (以下、非エキスパート

被験者) それぞれの正誤により 3 パターンに大別した 1on1 メモ (表 6) について詳しく調べた。

表 6: 各正誤パターン別の 1on1 メモの例

パターン	1on1 メモ文章
1 Expert: ○ Non-Expert: ○ BERT: ○	今の仕事が難しく、職場の中で無力感を感じており、日々の業務において気が重いとのこと。[不調]
2 Expert: ○ Non-Expert: × BERT: ○	新しく担当している業務に苦戦しているが少しずつ慣れてきているとのこと。業務負荷が高い割には問題ない様子であり、困っている様子も一切ないようである。[不調]
3 Expert: ○ Non-Expert: × BERT: ×	仕事で特に悩みはなく、自身としては職場の同僚ともコミュニケーションを以前よりうまく取れているとのことなので特段心配はない。[不調]

#### (1) パターン 1

パターン 1 は、BERT-finetuned、エキスパートである E1、非エキスパート被験者のすべてで不調と推定できた 1on1 メモを参考に著者が作成した例である。例に示す通り、このパターンの 1on1 メモでは「気が重い」など明らかにワークメンタリティが不調であることを示唆する表現が含まれているため、容易に推定できたと考えられる。

#### (2) パターン 2

パターン 2 は BERT-finetuned とエキスパートである E1 が正解し、非エキスパート被験者の 7/8 が不正解だった 1on1 メモを参考に著者が作成した例である。例では、「慣れてきている」や「問題ない」、「困っている様子も一切ない」などワークメンタリティが不調ではないことを示唆する表現が多いため、一見すると不調ではないように見えるという例である。このパターン 2 のメモについて、エキスパートである E1 にインタビューしたところ、不調判定した理由として「書き方として、メンバーが本当は不調であるにもかかわらず、上司が「特に問題ない」と言っているように感じた。」とのことであった。

このパターンにおいて、なぜ BERT-finetuned が正しく不調と判定できたのかを SHAP[Lundberg 2017] を用いて可視化した結果を図 1 に示す。図 1 において、赤字のトークンは BERT-finetuned において不調推定にプラスに貢献したものであり、青字は逆に好調推定にプラスに貢献したものである。また、色の濃さは貢献度の高さを示しており、色が濃いほど貢献度が高い。図 1 から「慣れ」や「問題ない」といったトークンでは青字になっていることから、BERT-finetuned においても確かにこういった表現については好調と推定されているのが分かる。一方で、「苦戦」「業務負荷」などのトークン付近で赤字になっているため、最終的には不調判定となっている。このことから、たとえ「問題ない」といった好調を示唆するポジティブな表現が含まれていたとしても、不調を示唆する表現がある場合には BERT-finetuned モデルでは不調であることを見逃さなかったと言える。

新しく担当している業務に苦戦しているが  
少しずつ慣れてきているとのこと。業務負  
荷が高い割には問題ない様子であり、困  
っている様子も一切ないようである。

図 1: パターン 2 の 1on1 メモの SHAP による可視化

### (3) パターン 3

パターン 3 は BERT-finetuned と非エキスパート被験者の 7/8 が不正解で、エキスパートである E1 のみが正解した 1on1 メモを参考に著者が作成した例である。このパターンにおいて、なぜ BERT-finetuned が誤って好調と判定してしまったのかを SHAP を用いて可視化した結果を図 2 に示す。図 2 からほとんどのトークンで青字になっており、かつ濃い赤字もないことから、全体的に好調な 1on1 メモとして判定されたことが分かる。

このパターン 3 のメモについて、エキスパートである E1 にインタビューしたところ、不調判定した理由として「“以前よりも”や“特段”といった表現から、メンバーが上司に対して心配させないように大丈夫だと言っているように見えた。」とのことであった。以上のことから、BERT-finetuned においては、パターン 3 のように不調を示唆する表現が含まれないもしくは少ない場合には不調と判定することは難しいと考えられる。

仕事で特に悩みはなく、自身としては職場の同僚ともコミュニケーションを以前よりもうまく取れているとのことなので特段心配はない。

図 2: パターン 3 の 1on1 メモの SHAP による可視化

### 4.3 考察

人による推定精度との比較の結果、BERT-finetuned で非専門家よりも高精度に推定できることが分かった。パターン 2 で示したように、好調を示唆する表現にひきずられることなく、不調を示唆する表現を見逃さなかったことが非専門家よりも高精度で推定できた理由と考えられる。

一方で専門家の推定精度と比べると同等からやや劣る結果となった。パターン 3 で示したような例では、パターン 2 に比べて不調を示唆する表現が含まれていないため、専門家のように特定の表現から不調を判定することができなかったと考えられる。よって更なる精度向上のためには、こういったパターンの 1on1 メモを多く学習させるなどの手法が必要である。

また、機械学習モデルと人の推定精度の比較結果以外にも、人の推定精度において非専門家のアンサンブル(平均)が、非専門家の中のベストメンバー(S7)に次ぐ精度を示した点は興味深い。このことは、複数の目で判定することが平均的には精度の向上につながる可能性と、しかしそれでも専門家や BERT モデルには劣るという限界を同時に示していると考えられる。

## 5. まとめと今後の展望

本研究では、上司の記録した 1on1 メモを用いてメンバーのワークメンタリティを推定する機械学習モデルを複数作成し、推定精度を検証した。その結果、1on1 メモによるファインチューニングした BERT モデルが最も高精度で F1-Score は 0.69 であった。

さらに人による推定実験を実施し、人による推定精度をベースラインとして比較した結果、1on1 に精通するエキスパートと同等の精度で分類できることが明らかになった。以上のことから、1on1 メモからそのメモが書かれた付近のメンバーの好不調を推定しうることを示した。

今回の検証では 1on1 メモが書かれた付近のワークメンタリティの推定を行ったが、今後は将来のワークメンタリティの予測についても検証していきたい。例えば半年後のワークメンタリティを目的変数とし学習することで、1on1 メモが記録された当時は

不調でなくても半年後などの将来的に不調になるメンバーを予測できる可能性がある。これにより不調になってからサポートするのではなく、早めに不調になりそうな兆候を検知し、サポートできるため今回の推定よりも実務的な価値が高いものになると考えられる。

### 参考文献

- [Cropanzano 2017] Cropanzano, R., Dasborough, M. T., & Weiss, H. M. "Affective events and the development of leader-member exchange." *Academy of Management Review*, 42, 233-258, 2017.
- [Weiss 1996] Weiss, H. M., & Cropanzano, R. "Affective events theory." *Research in organizational behavior*, 18(1), 1-74, 1996.
- [リクルートマネジメントソリューションズ 2022] リクルートマネジメントソリューションズ. "1on1 ミーティングに関する実態調査" 2022 年, [https://www.recruit-ms.co.jp/issue/inquiry\\_report/0000001055/](https://www.recruit-ms.co.jp/issue/inquiry_report/0000001055/)
- [Devlin 2018] Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805*, 2018.
- [Le 2014] Le, Quoc, and Tomas Mikolov. "Distributed representations of sentences and documents." *International conference on machine learning*. PMLR, 2014.
- [Chen 2016] Chen, Tianqi, and Carlos Guestrin. "Xgboost: A scalable tree boosting system." *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 2016.
- [Lundberg 2017] Lundberg, Scott M., and Su-In Lee. "A unified approach to interpreting model predictions." *Advances in neural information processing systems* 30 (2017).