QFセミナー

機械学習を利用した金融専門極性辞書構築手法・利用手法に関する研究

有村 亮一

2023年3月11日

東京都立大学大学院経営学研究科（博士前期課程）経営学修了

目次

研究の背景と概要

研究の背景と概要

本研究では金融専門極性辞書の元となる種辞書自動作成手法及びその発展的手法について取り扱う．実証研究もそれぞれ実施.

- テキストデータの重要性の増大
  - テキストデータに埋め込まれた情報，それを解析するインフラや技術が整備・増大
  - ファイナンスの分野でも実証研究を含めた各種研究に利用されている.
- 専門極性辞書の有無によって，分析結果が大きく変わる
  - 一般的な極性辞書にネガティブと分類されている単語のうち，75%はネガティブと分類されない (Loughran and McDonals(2011))
- 金融専門極性辞書の構築・構築手法の改善
  - 極性辞書とは単語に素性・性質が与えられたもの．直接的に点数が与えられることもある.

先行研究

先行研究：極性辞書

- 専門家による手作成
  - Loughran and McDonald（2011）: 過去の 14 年分の 10-K 資料から一定のルールでワードリストを抽出．その後，単語を negative, positive, uncertainty 等の極性に分類．
- 種辞書を利用するもの (極性タグ要)
  - Ito *et al.*(2020)　高次元ベクトル空間への単語埋め込み実施．コサイン類似に基づいた単語クラスタリングを実施．手作成の辞書を種辞書として，ニューラルネットワークの重みとしてセットし，文書の極性判定を学習する．クラスタから単語に極性が伝播する．
- 種辞書を不要とするもの（極性タグ不要）
  - 五島・高橋 (2017) の例: 株価の上昇，下落（異常リターンを標準化）を利用した極性辞書の作成 (教師スコアの自動付与)．データは，ニュース及びそのメタデータであるキーワードの単語頻度ベクトルを利用．

本研究で利用するニューラル・ネットワーク:GINN(Gradient Interpretable Neural Networks)

GINN モデル: 構築プロセス概略

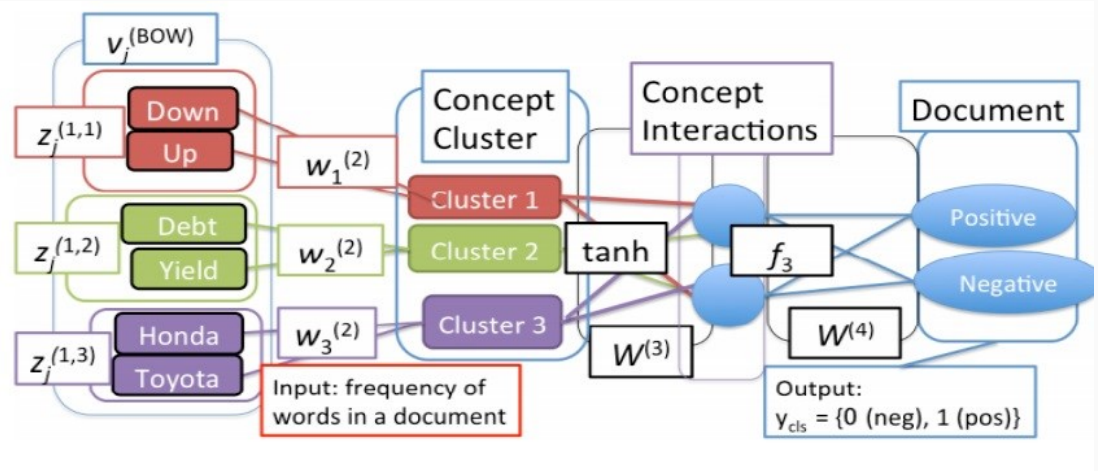
本研究では Ito *et al.* (2020) で取り上げられた GINN:Gradient Interpretable Neural Networks を取り上げて，極性辞書作成に利用する．構築手法は以下の通り．特に GINN 固有なものは種辞書の点数を初期値として入力層と第 2 層の間の重みに登録する Init 及び第 1 層及び第 2 層間のアルゴリズム (Algorithm 1:Update strategy) の記述.

1. 極性タグ付文書の文書を形態素解析し，単語に分割．単語を Word2Vec で高次元ベクトルに埋め込む.
2. 単語ベクトルを利用し Spherical k-means 手法で，任意の数のクラスタに分割する.
3. 種辞書を専門家６名で作成する（重要な単語[1]に各人員が点数をつけ，平均をとる）
4. Init: 点数を有する単語のみ，当該点数を入力層と第 2 層の重みに登録．点数がない単語は 0 と登録.
5. 誤差逆伝播時における第 2 層と第 1 層の間にクラスタの情報を利用する特有のアルゴリズム (Algorithm 1:Update strategy) を記述.

尚，本研究で言う種辞書とは入力層と第 2 層間の重みに登録する初期値，
金融極性辞書とは学習終了後（重み更新後）の入力層と第 2 層間の重みのことである.

[1] 重要な単語 (”important words”) 選定基準に関する記述は Ito *et al.*(2020) には見当たらない.

GINN モデル: 概要及び図解



- 文書のポジネガを予測（2 クラス分類）
- ウェイトとバイアスを誤差逆伝播法で調整し，予測誤差を最小化する．
- 全部で四層．活性化関数は全ての層で tanh を利用．
  - 入力層: 単語頻度ベクトル (但しクラスタ毎．クラスタについては次頁参照)
  - 第 2 層: ニューロン数=クラスタ数.
  - 第 3 層: ニューロン数任意 (上表では 2)
  - 出力層: ポジティブかネガティブの 2 クラスの内，1 つのクラスを選択

種辞書作成手法提案

種辞書作成プロセス: 単語極性評価関数 pol の定義

$$\mathbb{J} := \text{文書全体の集合(将来の文書も含む)} \tag{1}$$

$$\mathrm{Id} := \text{観測時点における文書のユニーク}ID\text{の集合} \tag{2}$$

$$\mathbf{j} : \mathrm{Id} \to \mathbb{J} \tag{3}$$

$$\mathrm{Id}^S := \mathrm{Id}^S \subset \mathrm{Id}\ \text{種辞書を作成する為の文書}ID\text{の集合} \tag{4}$$

$$\mathbb{p} : \mathrm{Id}^S \to \{-1, 0, 1\} \tag{5}$$

$$\mathbb{W} := \text{単語集合全体(未来の単語含む)} \tag{6}$$

$$\mathrm{freq}_j(w) := \text{文書}j\text{における単語}w\text{の頻度} \tag{7}$$

$$\mathrm{pol}(w; Id^S) := \sum_{i \in \mathrm{Id}^S} \mathrm{freq}_{\mathbf{j}(i)}(w)\mathbb{p}(i) \tag{8}$$

種辞書作成プロセス: 単語極性評価関数 pol の重みへの変換

$$W \subset \mathbb{W},\ W = \{w_1, w_2, \ldots, w_{\#W}\} \tag{9}$$

$$\mathbf{w} := [\tanh(\mathrm{pol}(w_1; \mathrm{Id}^{\mathrm{s}})), \ldots, \tanh(\mathrm{pol}(w_{\#W}; \mathrm{Id}^{\mathrm{s}}))]^{\top} \tag{10}$$

- (10) 式で tanh を利用することにより (-1,+1) の範囲に転換する.
- (10) 式の列ベクトル $\mathbf{w}$ の各要素こそが GINN の第 1 層と第 2 層間に登録する単語の重みの初期値である.

実証研究

利用データ:Yahoo!Japan ファイナンス掲示板からのデータ

- 2022 年 1 月 1 日から 2022 年 7 月 9 日まで
- 全 33 セクターに属する 271 社の全ての掲示板が対象

sentiment	date	comment	user	sector	mb_title
強く買いたい	2022/07/09 7:11	しんさん夜のお勤めご苦労様でした！まろんさん久々です。マサボンさんちょっとずつ加勢させてもらいますよ！とりあ	5c4cd95be6f	サービス業	7078 - ＩＮＣＬＵＳＩＶＥ(株) 2022/07/06～
強く買いたい	2022/07/09 6:44	あんまり糞糞言うと自分が糞になるよ(^.^)	3c685b608fd	サービス業	6081 - アライドアーキテクツ(株) 2022/05/14～
強く売りたい	2022/07/09 0:34	わしはカスブンは、喋る相手もいなく　掲示板だけが友達です！！！勃起ですか？童貞ですが人一倍性欲が強いです、	9db882921b(	サービス業	4347 - ブロードメディア(株) 2022/07/03～
強く売りたい	2022/07/08 22:51	まっ！月曜日昼から878だからな！グロース舐めんなよ	ec0da24fe90	サービス業	7359 - (株)東京通信 2022/07/01～2022/07/08
強く買いたい	2022/07/08 22:30	そうなの？　むしろ　全国の警備の　みなおしと　強化やろ？	39d4482060:	サービス業	9735 - セコム(株) 2021/06/08～
強く買いたい	2022/07/08 22:27	とりあえず夜PTSで少し購入。２９００円まで平均さげましたので、週明け逃げられるでしょう。さすがに初日の揺り戻し	ee1785e3a8:	サービス業	9556 - ＩＮＴＬＯＯＰ(株)
強く売りたい	2022/07/08 21:18	これ犯行声明みたい　なんかムカつく	e060507462(	サービス業	4347 - ブロードメディア(株) 2022/07/03～
強く売りたい	2022/07/08 21:01	ちがうちがう　かもめが変わりは無理　１００００%ムリただ撃たれるの　変わって欲しい　純粋に　(^^♪	e060507462(	サービス業	4347 - ブロードメディア(株) 2022/07/03～
強く買いたい	2022/07/08 20:45	例え空き地になっても土地さえ持ってたらお宝NFTが埋まっているかも見たいな宝探し見たいなお楽しみ要素希望黄金(	d46d47114c4	サービス業	2437 - ＳｈｉｎｗａＷｉｓｅＨｏｌｄｉｎｇｓ(株) 2022/07/08～
強く買いたい	2022/07/08 20:34	いつ位にアンバサダーを発表するとでも言ってくれたらそれに向けて上がるだろうけど名前が出る前に皆んな一斉に売り	d46d47114c4	サービス業	2437 - ＳｈｉｎｗａＷｉｓｅＨｏｌｄｉｎｇｓ(株) 2022/07/08～
強く買いたい	2022/07/08 18:09	最近調子が良いので買い増し	2e3affb8388	サービス業	9625 - (株)セレスポ 2022/04/19～
強く買いたい	2022/07/08 16:07	下げすぎー。	9451fafc7ad(	サービス業	6552 - (株)ＧａｍｅＷｉｔｈ 2022/07/07～

GINN での実証研究：モデル設定

Simple-GINN

- 損失関数:binary crossentropy
- クラスタ数 1
- 第 3 層ニューロン数 5

GINN モデル

- 損失関数:binary crossentropy
- クラスタ数 10
- 第 3 層ニューロン数 5

尚，データの数 n はそれぞれ約 250 行，約 500 行で検証した．評価指標についてはマクロ F1 スコア（付録 1 参照）を利用した.

実証研究: 結果詳細 (データの数約 250)

Table 1: Simple-GINN による学習及び予測結果

学習データ数	単語数	学習: 対象セクター	予測（マクロ F1 スコア[2]）	学習時間 [3]（概算）
252	7, 712	全セクター	0.225	5 時間 10 分
253	7, 532	情報・通信セクター	0.761	5 時間 00 分
254	6, 281	食料品セクター	0.865	3 時間 30 分
239	6, 583	銀行セクター	0.133	3 時間 40 分

Table 2: GINN による学習及び予測

学習データ数	単語数	学習: 対象セクター	予測（マクロ F1 スコア）	学習時間（概算）
250	7, 265	全セクター	0.241	4 時間 50 分
253	7, 532	情報・通信セクター	0.682	7 時間 30 分[4]
254	6, 281	食料品セクター	0.151	3 時間 40 分
239	6, 583	銀行セクター	0.165	3 時間 45 分

[2] 学習は東京都立大学経営学研究科管理の計算サーバー Turing で実施

[3] マクロ F1 スコアについては後添

[4] 学習した計算機の計算能力が Turing より劣後する為，参考値.

実証研究: 結果詳細 (データの数約 500)

Table 3: Simple-GINN による学習及び予測結果

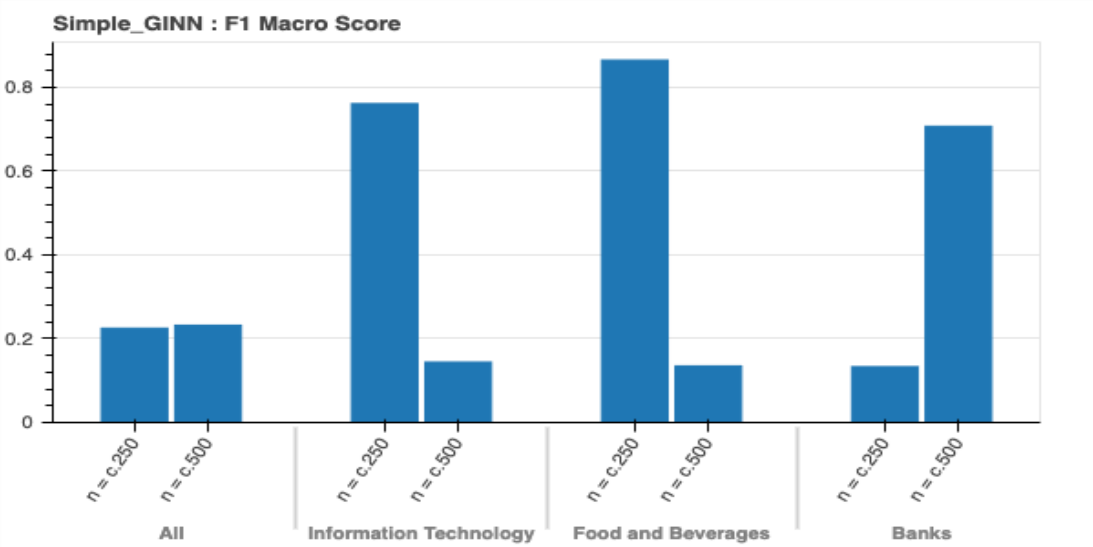
学習データ数	単語数	学習: 対象セクター	予測（マクロ F1 スコア）	学習時間[5]（概算）
493	約11,000	全セクター	0.232	0 時間 35 分
497	約10,300	情報・通信セクター	0.144	0 時間 34 分
503	約11,000	食料品セクター	0.707	0 時間 34 分
469	約9,300	銀行セクター	0.135	0 時間 32 分

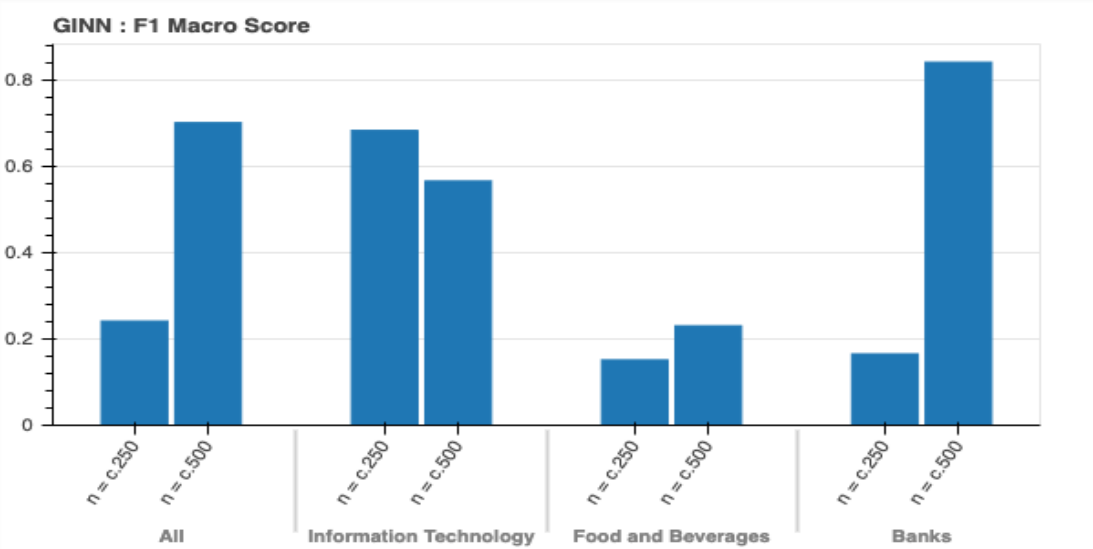
Table 4: GINN による学習及び予測

学習データ数	単語数	学: 対象セクター	予測（マクロ F1 スコア）	学習時間[5]（概算）
493	約11,000	全セクター	0.70	5 時間 00 分
497	約10,300	情報・通信セクター	0.56	5 時間 00 分
503	約11,000	食料品セクター	0.23	4 時間 00 分
471	約9,300	銀行セクター	0.84	5 時間 00 分

[5] 学習は個人用のローカル端末で行なった．前項で実施した学習時間より一部短くなっているのはアルゴリズムを変更したことによるもの．

実証研究：結果サマリ


Simple_GINN : F1 Macro Score
0.8
0.6
0.4
0.2
0
n = c.250
n = c.500
All
Information Technology
Food and Beverages
Banks


GINN : F1 Macro Score
0.8
0.6
0.4
0.2
0
n = c.250
n = c.500
All
Information Technology
Food and Beverages
Banks

文書ストリーム

時間概念の導入

過去に観測された単語は現在の単語より重要度が低いことを仮定し，(8) 式に時間概念を導入するもの.

$$\mathbf{t} : \mathrm{Id}^{\mathrm{S}} \to [0, \infty) \text{ 文書 ID の観測時刻を返す関数} \tag{11}$$

$$\mathbf{t}(i) := \mathrm{ID}i\text{を持つ文書の観測時刻} \tag{12}$$

辞書の作成時刻 $t_0 \in [0, \infty)$ での単語 $w \in \mathbb{W}$ の評価 pol_{t_0} は以下で与えられる.

$$\mathrm{pol}_{t_0}(w; \mathrm{Id}^{\mathrm{s}}) := \sum_{i \in \mathrm{Id}^{\mathrm{s}}} \mathbf{1}_{\{\mathbf{t}(i) \le t_0\}} e^{-\lambda(t_0 - \mathbf{t}(i))} \mathrm{freq}_{\mathbf{j}(i)}(w) \mathbb{p}(i) \tag{13}$$

但し，$\lambda \ge 0$ は減衰定数である.
また，辞書の作成時刻 t_0 を動かすことで同じ学習データでも複数の辞書の版 (エディション) を作成することができる.

実証実験: 時間減衰モデル設定

GINN 時間減衰モデル[1]

- 損失関数:binary crossentropy
- クラスタ数 10
- 第 3 層ニューロン数 5
- $\lambda = 1.38$

[1] 今回の実験では，関数 $\mathbf{t}$ の実装は，実時間ではなく，以下のような手法を用いて，各観測文書に時刻 $t \in [0, 1]$ をアサインした．すなわち，データセット Id^S を実時間の順に $i_0, i_1, \cdots, i_{N-1}$ と並べた後，$\mathbf{t}(i_u) := \frac{u}{N}$ と時刻をアサインした．そのうえで，λ の値としては，$\frac{N}{2}$ 個前に観測した文書の影響を半減するようにセットした．すなわち，$e^{-\lambda \times 0.5} = 0.5$ を解いて，$\lambda = 1.38$ とした．

実証実験: 結果詳細 (データの数=約 250)

Table 5: GINN による学習及び予測 (再掲)

学習データ数	単語数	学習: 対象セクター	予測（マクロ F1 スコア）	学習時間（概算）
250	7,265	全セクター	0.241	4 時間 50 分
253	7,532	情報・通信セクター	0.682	7 時間 30 分
254	6,281	食料品セクター	0.151	3 時間 40 分
239	6,583	銀行セクター	0.165	3 時間 45 分

Table 6: GINN 時間減衰モデル: $\lambda = 1.38$

学習データ数	単語数	学習: 対象セクター	予測（マクロ F1 スコア）	学習時間（概算）
252	7,712	全セクター	0.202	5 時間 00 分
253	7,532	情報・通信セクター	0.188	5 時間 00 分
254	6,281	食料品セクター	0.200	3 時間 40 分
239	6,583	銀行セクター	0.730	3 時間 46 分

実証実験: 結果詳細 (データの数=約 500)

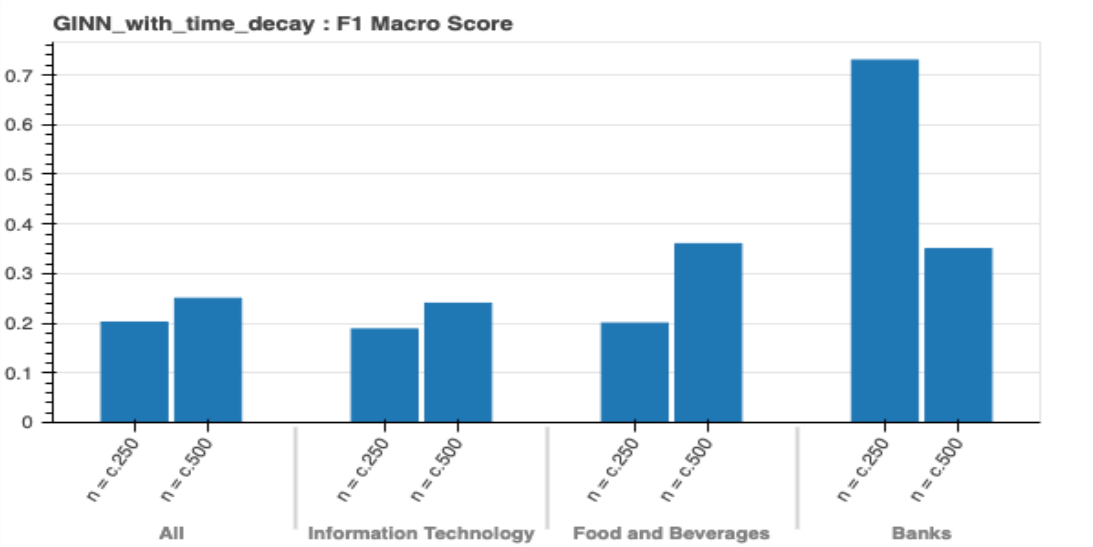
Table 7: GINN による学習及び予測 (再掲)

学習データ数	単語数	学習: 対象セクター	予測（マクロ F1 スコア）	学習時間（概算）
493	約11,000	全セクター	0.70	5 時間 00 分
497	約10,300	情報・通信セクター	0.56	5 時間 00 分
503	約11,000	食料品セクター	0.84	4 時間 00 分
471	約9,300	銀行セクター	0.23	5 時間 00 分

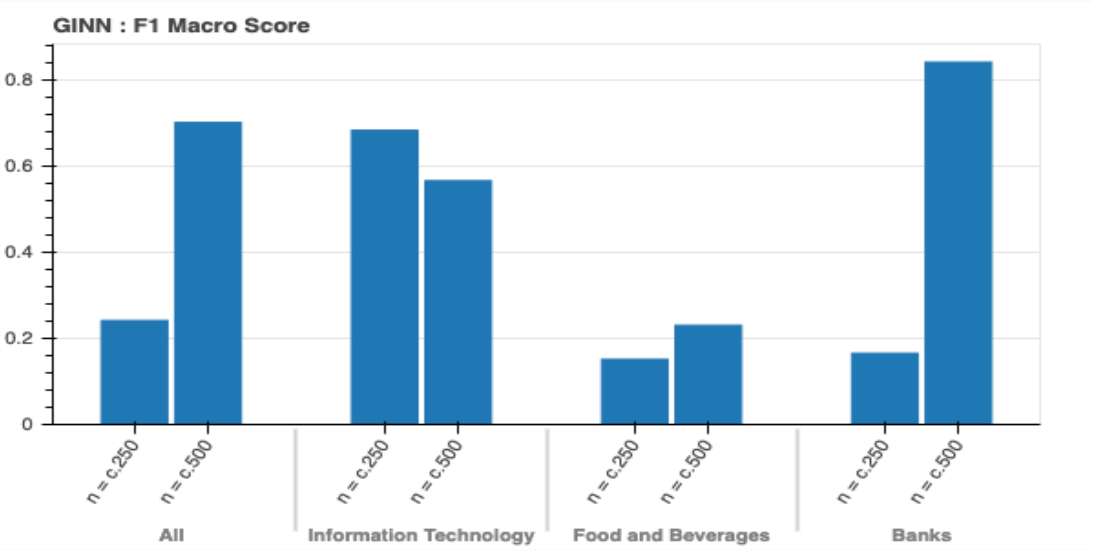
Table 8: GINN 時間減衰モデル: $\lambda = 1.38$

学習データ数	単語数	学習: 対象セクター	予測（マクロ F1 スコア）	学習時間（概算）
493	約11,000	全セクター	0.25	5 時間 50 分
497	約10,300	情報・通信セクター	0.24	5 時間 00 分
503	約11,000	食料品セクター	0.35	7 時間 00 分
470	約9,300	銀行セクター	0.36	3 時間 00 分

実証研究：結果サマリ



下図再掲



結論: 本研究の成果及び今後の課題

結論

本研究の成果：

- 種辞書自動作成の手法の確立
  - セクター毎の辞書を容易に作ることができるようになった.
  - 時間概念を導入しタイムディケイを極性に組み込むことが可能となった.
- GINN の実装
  - algorithm 1 : Update Strategy の実装
  - 入力ベクトルと重みベクトル (行列ではない) の内積計算の実装

時間概念を極性辞書構築に組み込んだ研究は見当たらず，今後もその有効性検証とともに，複数タイミングでの辞書生成，市場のレジームチェンジ時の極性判定において，着目すべき単語を時系列の情報を利用して作成する手法，複数のエディションの利用など種々の発展方法が適用可能と思われる.

結論

今後の課題:

- 学習時間の短縮/計算速度の改善
  - アルゴリズムの見直し (e.g., 貪欲計算からグラフ計算への移行検討)
- 実証実験の拡充
  - 重みの初期値にランダムを与えた場合と種辞書を利用した場合比較
  - エディションを利用した特定の期間の辞書の有効性検証
  - 文書ストリームとしてデータをリアルタイム予測
  - 全個別セクターの種辞書作成及び比較
  - 個別企業の種辞書作成及び比較
  - クロスバリデーションの実施によるハイパーパラメーターの調整
  - 学習データを増加させた場合の影響の確認 (250 千件のデータのうち利用しているのは 0.1%〜0.2%程度のみ)
  - より長い時系列において単語のタイムディケイの半減期を長めに取り，市場のレジーム・チェンジ時における辞書の有効性を検証
  - データセットの変更 (Yahoo!掲示板から QUICK で提供されいている金融情報ニュースデータに変更)
  - 極性評価関数 *pol* の変更
  - 頻度を重みに変更する際の *tanh* 以外の関数の利用の検討

参考文献

参考文献

Ito, T., Sakaji, H., Izumi, K., Tsubouchi, K., and Yamashita, T. (2020) “GINN: gradient interpretable neural networks for visualizing financial texts,” *International Journal of Data Science and Analytics*, **9**(4), 431–445.

Loughran, T. and McDonald, B. (2011) “When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks,” *The Journal of Finance*, **66**(1), 35–65.

五島圭一・高橋大志 (2017)「株式価格情報を用いた金融極性辞書の作成」,『自然言語処理』, **24**(4), 547–577.

付録

付録 1: マクロ F1 スコアについて

本研究で利用しているマクロ F1 スコアとは，各クラス毎の予測について下記 F_1 スコアを求め，クラス数（本実証研究の場合はポジティブクラスとネガティブクラスの 2) で割ったもの．多クラス分類問題における分類器（モデル）の性能判定に使われ，$[0, 1]$ の範囲をとる．数字が大きい方が性能が高い．

$$F_1 Score := \frac{2 \times recall \times precision}{recall + precision} = \frac{2}{\frac{1}{recall} + \frac{1}{precision}} \tag{14}$$

$$recall = \frac{True\ Positvie}{True\ Positive + False\ Negative} \tag{15}$$

$$precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \tag{16}$$

付録 2:GINN のコード実装について

本論で利用した GINN モデルの python コードについては，github に掲載している[6]．尚，執筆時点では完成物のモデルについて，技術的な理由で main ブランチへマージを行っていない．別ブランチである batch-dimension-fix[7]の ginn_model.py が完成物であるので，参照されるときはこちらをご参照されたい．将来的には main ブランチに統合・整備する予定である．

[6] https://github.com/r-arimura7/GINN

[7] https://github.com/r-arimura7/GINN/tree/batch-dimension-fix

付録 3: 先行研究: 金融専門極性辞書 手法プロ・コン

- 専門家による手作成
  - プロ: 正確
  - コン：作成コスト高い
- 種辞書を利用するもの
  - プロ：作成コストが手作成辞書よりも低い
  - コン：教師データが少量ではあるが必要となる
- 種辞書を不要とするもの
  - プロ: 教師データが必要ない，作成コストが低い
  - コン: 教師データはいらないものの，利用できるデータについて制限があると思われる（より多くのメタ・データが必要など）