Sentiment analysisを用いた
SNS分析の活用例

social Media Analysis for Product Safety using Text Mining and Sentiment Analysis
ソーシャルメディア分析 テキストマイニングとセンチメント分析を利用した製品安全性の向上

（2015年10月18日に提出）Haruna Isah, Daniel Neagu, Paul Trundle

https://arxiv.org/abs/1510.05301

どんなもの？

ソーシャルメディアのデータから、機械学習技術を用いて医薬品や化粧品のユーザーの見解や経験を示唆する感情を推論するために使用できるかを実証している。

➡831個のポジティブメッセージのうち300個(36%)がネガティブに正しく分類されているのに対し、2027個のネガティブメッセージのうち180個(8.9%)がポジティブに正しく分類されており、合計の精度は約83%だった。

どうやって有効だと検証した？

・テキストの収集・・・TwitterとFacebookのAPIを使用
・前処理・・・文章を構造化し大規模に集積したもの（コーパス）を特徴ベクトルに変換、トークン化。逆文書頻度を求め、単語毎の重要度を算出。
・センチメント分析・・・分類器としてNaive Bayes を使用し、ツイートやコメントを適切なセンチメントクラスに分類する。

技術の手法や肝は？

・IDF(Inverse Document Frequency)・・・文書集合の中のある単語が含まれる文書の割合の逆数を表す．単語が他の文章にも多く出現しているほどIDF値は小さくなり，単語が他の文章にあまり出現していないほどIDF値は大きくなる
・intuitionistic fuzzy reasoning・・・メンバーシップ関数、非メンバーシップ関数、およびhesitant 関数を使用して、サンプルトレーニングによって機能の不確実性を定量的に表すとともに、程度の副詞、接続詞、否定的な単語の影響を受ける感情表現を考慮している。そして、特徴の直観的なファジィ情報の集合体を用いて、フレーズ、文、テキストの順序のレベルで、テキストの意味的方向性を合成する。

議論はある？

特に無し

先行研究と比べて何がすごい？

Naive-Bayes 分類器の欠点として報告されているのは、特徴が互いに独立しているという仮定であり、マキシマムエントロピーは疎なデータの場合にオーバーフィットに悩まされていた。ソーシャルメディアプラットフォーム上のステータスアップデート、ツイートやコメントとして報告された人気のある医薬品や化粧品のブランドの消費者の意見や経験を活用しファジーセンチメントスコアリングのような計算知能技術の組み合わせた。

次に読むべき論文は？

X.F. Li　D. Li(2013)Sentiment orientation classification of webpage online commentary based on intuitionistic fuzzy reasoning
https://www.researchgate.net/publication/289977741_Sentiment_orientation_classification_of_webpage_online_commentary_based_on_intuitionistic_fuzzy_reasoning

日付

Sentiment Analysis Using Simplified Long Short - term Memory Recurrent Neural Networks
簡明化したセンチメント分析を利用した 長期短期記憶リカレントニューラルネットワーク

（2020年5月8日に提出]）Karthik Gopalakrishnan, Fathi M.Salem

https://arxiv.org/pdf/2005.03993.pdf

どんなもの？

・GOP Debate Twitterのデータセットを用いてセンチメント分析を行う(2016年大統領指名のためにオハイオ州で行われた8月初旬のGOP討論会に関する数万件のツイート)
その際、3つの異なるLSTMそれぞれを使った場合でaccuracyを算出している
➡全てのモデルで、ポジティブな感情よりもネガティブな感情を正確に予測した。

どうやって有効だと検証した？

①前処理：データセット内でラベル付けされたポジティブ、ネガティブ群のみを抽出した。そしてすべての単語を下位フォーマットに変換しすべての文を数列に変換した（文字と数字のみに限定し、特徴量を制限
②NNは11層からなり、spatial dropout 1D層,1dCNN層とmax_pooling層を通して特徴量を抽出。そして3つの異なるLSTMを中間層に置き換えたRNNを通して最後に、ドロップアウトを伴う3つの連続した密な層を用いている。

技術の手法や肝は？

学習を高速化し、計算コストと時間を削減するために、LSTMモデルの6つの異なるパラメータを削減したスリム版(slim LSTM)を提案した。

議論はある？

ポジティブな感情よりネガティブな感情を正確に予測したのは、訓練データセットのバランスが悪いためだと考えられる。（ポジティブ・ネガティブが偏っていた）
訓練データセットを人為的にバランスをとることで、ポジティブセンチメント分類を改善し、ネットワークの全体的な性能を向上させることができると考えられる。

先行研究と比べて何がすごい？

標準的なLSTMネットワークと、２種類の削減されたSlim LSTMモデルの性能を比較している。また、双方向LSTM層の導入が性能に与える影響についても検討している。

次に読むべき論文は？

Hochreiter, Sepp & Schmidhuber, Jürgen. (1997).
LongShort-term Memory. Neural computation. 9. 1735-80.10.1162/neco.1997.9.8.1735
https://www.bioinf.jku.at/publications/older/2604.pdf

2020/6/18

Heterogeneous Graph Attention Networks for Early Detection of Rumors on Twitter
異種グラフ注意ネットワーク ツイッターでのうわさの早期発見のために

（2020年6月10日に提出]）Qi Huang、Junshuai Yu、Jia Wu、Bin Wang

https://arxiv.org/pdf/2006.05866.pdf

どんなもの？

ツイッター上で広がっている噂を早期発見する。
ツイートをツイートとユーザー情報（tweet-user）、ツイートとツイート内の単語（tweet-word）に情報を分解しHANを用いて分析した。
➡噂のテキスト内容を含む分解されたtweet-wordサブグラフの方が、噂の発信元ツイートの伝播を含む分解されたtweet-userサブグラフよりも、噂の検出に大きな効果があることがわかった。

どうやって有効だと検証した？

データ・・・公開されている2つのTwitterデータセット。非噂、偽の噂、真の噂、または検証されていない噂としてラベル付けされている
噂の早期発見のプロセスをシミュレートするために、噂の発信元ツイートが投稿されてからの経過時間やユーザーのリツイート数を制御して、噂の伝播の異なる期間を表現し、異なる期間における噂の検出精度を計算して性能を評価。

技術の手法や肝は？

Heterogeneous Graph Attention Network(HAN)
・・・異なる種類のデータで構成されていグラフデータに対する教師あり学習。
グラフデータに対してディープラーニングを適用しようとする、いわゆるGraph Neural Networks（GNN）を、異なる種類のデータで構成されているグラフを対象としたタスクに取り組むことができるようにしたもの

議論はある？

特に無し

先行研究と比べて何がすごい？

既存の噂検出のための手法の大部分は、テキストコンテンツ、ユーザープロファイルなどから識別された特徴を抽出するために、特徴工学を利用していた。しかしこれらの方法では、テキスト内容の様々な領域にわたる意味的関係を十分に利用できていない。
本研究では、テキストコンテンツの異なる種類のグラフの意味関係を、HANを用いて発信元のツイート伝播のグローバルな構造情報とともに捉えられるようになった。

次に読むべき論文は？

Xiao Wang, Houye Ji, Chuan Shi, Bai Wang, Peng Cui, P. Yu, Yanfang Ye　”Heterogeneous Graph Attention Network”
https://arxiv.org/pdf/1903.07293.pdf

J. Ma, W. Gao, P. Mitra, S. Kwon, B. J. Jansen, K.-F. Wong, and M. Cha, “Detecting rumors from microblogs with recurrent neural networks.” in IJCAI, 2016, pp. 3818–3824.
https://www.ijcai.org/Proceedings/16/Papers/537.pdf

日付

Unsupervised machine learning to analyse city logistics through Twitter
Twitterを介して都市の物流を分析するための教師なし機械学習

（2019年6月18日に提出）Simon Tamayo (CAOR), François Combes (IFSTTAR/AME/SPLOTT), Gaudron Arthur (CAOR)

https://arxiv.org/abs/1906.07529

どんなもの？

Twitterを利用してシティロジスティクスに関するソーシャルメディアマイニングを行った。主に以下の2つの検討を行っている。(1) シティロジスティクスについてツイートした人が注目する概念を（出現頻度の点で）重要度が高いか低いかの概念を可視化し、それらの概念間の近接性を示す。(2) 収集したデータに対してセンチメント分析を行った。コーパスのセンチメント分布が中立 48%、肯定 45%、否定 7%であることから、シティ・ロジスティックスの全体的な見方が否定的よりも肯定的であることを評価することができた。コーパスの中で最もよく使われている n-gram を統計的に分析した結果、この分析では、シティロジスティクスのツイートの中で最も重要なトピックは雇用であることも示された。インタレストマップを見ると、雇用（求人）、新技術（自動運転車、ブロックチェーン、IoT）、スタートアップや新しい組織形態（ライドヘイル、宅配物流、ハイパーローカル物流）などの特徴的なクラスターが見えてきた

どうやって有効だと検証した？

データ収集は、検索用語「City Logistics」、「Last Mile Logistics」、「Urban Logistics」、「Urban Freight」でTwitterのウェブサイトをスクレイピングすることによって行われた。
・前処理・・・入力内容を特徴ベクトルに変換。次元削減(SVD)
・分析・・・K-Meansアルゴリズムをデータ(3)に適用。マニホールド学習アルゴリズムを適用し、2次元の結果を得た。使用したアルゴリズムはt-SNEで、複数の異なるマニホールドやクラスタに存在するデータを明らかにし、出力されたデータにNLTKのVADERを用いてセンチメント分析を行った。

技術の手法や肝は？

次元削減,
クラスタリング
t-SNE
VARDER・・・pythonライブラリのNLTKにおけるSentiment Analysis
マニホールド学習・・・多様体学習（二次元や三次元の形でデータを可視化できるようにデータの次元を縮約する）

議論はある？

Twitterでのコミュニケーションは、これらの目的（市場のストレスを分析したり、将来の発展を予測したり）に効率的に貢献していると思われる。ちなみに、研究者や政策立案者にとって、ソーシャルメディア・マイニングは、特にこのような目まぐるしく変化する 環境において、非常に費用対効果の高いビジネス・インテリジェンス・プロセスになり得る。ソーシャルメディア・マイニングがシティ・ロジスティクスの観察に何をもたらすかを評価するためには、十分に 表現されていない問題や盲点を特定することが重要である。

先行研究と比べて何がすごい？

従来シティロジスティクスに関してSNSデータを元に機械学習技術を用いて分析したことはなかった点。

次に読むべき論文は？

Olson, R.S. & Neal, Z.P., 2015. Navigating the massive world of reddit: using backbone networks to map user interests in social media. *PeerJ Computer Science*, 1, p.e4.

https://www.researchgate.net/publication/259288224_Navigating_the_massive_world_of_reddit_Using_backbone_networks_to_map_user_interests_in_social_media

日付

A Tweet-based Dataset for Company-Level Stock Return Prediction
企業レベルの株式リターン予測のためのツイートベースデータセット

（2020年6月17日に提出）Karolina Sowinska、Pranava Madhyastha

https://arxiv.org/abs/2006.09723

どんなもの？

ソーシャルメディアの投稿からの豊富な質的データを処理して以下の２点について検討している。

・ツイートに含まれる意味情報は、1日、2日、3日、7日の株式予測のベンチマーク精度を向上させるのか？
・多視点学習プロセスにおいて、複数種類の特徴を融合させることで、1日、2日、3日、7日の株式リターン予測の精度を向上させることができるか？

どうやって有効だと検証した？

データセット：サンプルとしてtwitterの財務情報について言及している文字情報、ラベルとして株式情報。
ツイート情報に関して3つのベンチマークを設計。ベンチマーク1は財務情報のみを含む。ベンチマーク2は、財務情報と「カウント」、例えば、リリースされたニュース記事の数や、特定の銘柄についてリリースされたツイートの数を含んでいる。最後に、ベンチマーク3は、独自リソースからのセンチメント情報を含む。
分析：自動機械学習ツール
評価指標：Accuracy

技術の手法や肝は？

Social Media Data Analyzer-Sentiment Analysis (SMeDA-SA)
Twitterの投稿に含まれるような曖昧な時間的データをマイニングする手法。
NLP技術を応用して、ツイートの感情を5つのカテゴリー（Positive+、Positive、Neutral、Negative、Negative-）に分類。また関連性発見アルゴリズムを使用して、ターゲットに関連するすべてのコンセプトを発見する（例えば、「MacBook Pro」は「Apple Inc.）次に概念マップと5つの感情カテゴリの間に関連性の度合いを算出する。最後に、別のアルゴリズムを適用して、株価とツイートから検出されたセンチメントの間に相関関係があるかを確認する。

議論はある？

Li et al.2017]では特定の企業のリターンの方向性を予測しているが、複数のデータソースを融合させず、Twitterのセンチメントのみを考慮していた。彼らの報告によると、3日間のリターンラベルが最も精度が高く、66.48%となっている。しかし今回の結果とこれらの研究で報告された結果を直接比較することはできない。これらの研究結果を引用して、我々の結果が似たような範囲にあることを説明するが、我々は異なるデータセットを使用しているため、これらの研究の実験結果は今回の研究と厳密に比較できるものではない。

先行研究と比べて何がすごい？

Li et al.2017ではツイートのセンチメント情報のみを考慮して特定の企業のリターンの方向性を予測していたが、本研究では財務情報に関するツイート内容・ツイート数・センチメント情報を考慮して総合的に検討している。

次に読むべき論文は？

Li, B., Chan, K. C., Ou, C., and Ruifeng, S. (2017). Dis- covering public sentiment in social media for predicting stock movement of publicly listed companies. *Informa- tion Systems*, 69:81–92.

https://www.sciencedirect.com/science/article/abs/pii/S0306437916304860