概要

感情分析は、テキストに表明されている書き手の意見を肯定的・否定的・中立のいずれかに分類するタスクであり、マーケティングや投資など多岐にわたる分野で活用されている。感情分析では、感情を表す単語やフレーズを収集し、それらの極性(肯定、否定、中立)のスコアを定義したデータベースである感情語辞書がよく利用されている。個々の単語が持つ極性スコアを基にテキスト全体の極性スコアを計算し、その結果が正であれば肯定的、負であれば否定的と判定する。しかし、単語の極性はテキストのドメインによって異なるため、汎用的な感情語辞書を用いると解析を誤る可能性がある。ドメインに固有の感情語辞書を用意することが望ましいが、感情語辞書を人手で構築するためには多くの時間と労力を要する。そのため、ドメインに固有のコーパスから高品質なドメイン固有の感情語辞書を自動獲得する手法が注目されている。

感情語辞書の自動獲得では、ドメインに特化した辞書を新たに作成する「辞書構築」(lexicon construction)と、既存の感情語辞書を特定のドメインに適応させる「辞書適応」(lexicon adaptation)の 2 種類のアプローチがある。辞書構築は、感情語を獲得するところから始め、当該ドメインにしか使われない感情語も含めて辞書を獲得する。一方、辞書適応は、新しい感情語は獲得しないが、感情語の極性やそのスコアを当該ドメインにあわせて修正する。

感情分析や感情語辞書の自動獲得に関する既存のサーベイでは、要素技術、データソース、対象言語などを軸に既存研究がまとめられてきた。しかし、感情語辞書の自動獲得のアプローチ(辞書構築もしくは辞書適応)と対象テキストのドメインの関係性に注目したサーベイは筆者の知る限り存在しない。

本課題研究では、手法のアプローチと対象テキストのドメインに着目して、感情語辞書の自動獲得に関する研究の動向を調査する。辞書の自動獲得のアプローチとして辞書構築と辞書適応の 2 つを考慮する。一方、ドメインとして、政治演説、ニュース、映画レビュー、商品レビュー、ソーシャルメディア、その他の 6 つを考慮する。Google Scholar を用いて論文を検索し、検索結果上位の論文の中から感情語辞書の自動獲得に関連する論文を人手で選別し、調査対象とする。辞書獲得のアプローチ×ドメインの行列を作成し、調査対象の論文のそれぞれをこの行列のセルに当てはめ、該当する論文が多いのはどのアプローチとドメインの組み合わせであるかといった全体的な研究動向を調査する。あわせて、個々の研究事例から、ドメインが辞書獲得のアプローチの選択に与える影響について考察する。

論文検索と人手による選別の結果、417 件の論文を調査対象とした。辞書獲得のアプローチとドメインのそれぞれの組み合わせについて、該当する論文数は

以下の通りである。辞書構築の論文については、(辞書構築,政治演説)は 3 件、(辞書構築,ニュース)は 26 件、(辞書構築,映画レビュー)は 33 件、(辞書構築,商品レビュー)は 112 件、(辞書構築,ソーシャルメディア)は 133 件、(辞書構築,その他)は 55 件であった。一方、辞書適応の論文については、(辞書適応,政治演説)は 2 件、(辞書適応,ニュース)は 2 件、(辞書適応,映画レビュー)は 11 件、(辞書適応,商品レビュー)は 25 件、(辞書適応,ソーシャルメディア)は 4 件、(辞書適応,その他)は 11 件であった。

辞書構築と辞書適応の論文数を比較すると、辞書構築の論文は全体の 86.8%、辞書適応の論文は 13.2%を占め、辞書構築の手法が主流であることがわかった。個々のドメインに着目すると、ソーシャルメディアドメインでは 97.1%、ニュースドメインでは 92.9%が辞書構築の論文であった。ソーシャルメディアやニュースでは新しい単語が使われることが多く、テキストから感情語を収集することから始める辞書構築の手法の方が適しているためと考えられる。一方、相対的に辞書適応の手応が採用されることが多かったドメインは、政治演説(40%)、映画レビュー(25.0%)、商品レビュー(18.2%)であった。新語が出現しにくい政治演説のドメインでは辞書適応の手法が適していると考えられる。映画・商品レビューでは、ドメイン固有の単語が使われるために辞書構築の手法が適していると予想していたが、実際には辞書適応の手法もある程度使用されている。その他のドメインに該当する研究では、動画コメント、ビデオゲーム、漢詩といったテキストを対象とし、特定のコンテキストやコミュニティに根ざした言葉や表現を分析するために、辞書構築のアプローチが積極的に採用されていることがわかった。これは、それぞれのドメインが持つ独自の言語、表現、文化的背景を正確に捉えるためには、ドメインに固有の感情語を自動的に収集する必要があるためと考えられる。

ドメインについて論文数を比較すると、商品レビュードメインとソーシャルメディアドメインの研究が最も多く、全体に占める割合はそれぞれ 32.9%であった。辞書構築の論文に限ると、ソーシャルメディアドメインの論文が 36.7%と最も多く、新語の多いソーシャルメディアでは辞書構築の手法が使われやすいという先の考察を裏付ける結果が得られた。ただし、ソーシャルメディアにおいても、新語を多く含む汎用の感情語辞書があれば辞書適応の手法も有望であることがわかった。一方、辞書適応の論文に限ると、商品レビュードメイン(45.5%)、映画レビュードメイン(20.0%)で用いられることが多く、レビューを対象とした感情分析には既存の汎用的な感情語辞書をレビューに適応させた辞書が使われることが相対的に多いことがわかった。

これらの結果から、感情語辞書の自動獲得アプローチは、その適用されるドメインに密接に関連していることが確認された。各ドメインの特性とニーズに応

じて最適なアプローチを選択し、適用することが、感情分析の精度と有効性を向上させる鍵であると言える。