悪口表現の検出手法について

Detection method of slander

東京工科大学大学院　バイオ・情報メディア研究科　コンピュータサイエンス専攻
思考と言語研究室　小林友則
指導教員　　亀田弘之、渡邊紀文、喜多義弘、相田紗織

キーワード：悪口，誹謗中傷，感情分析，Word2Vec，SNS

1. はじめに

近年、スマートフォンの普及が進み、多くの人がスマートフォンを所持している。青少年の SNS の利用者は年々増加しており、それに伴って SNS では悪口が蔓延している。悪口を目にしてしまう機会が増加し、心が傷ついてしまう、そのようなトラブルが今後も増加すると予想される。そこで機械のプログラムによる悪口のフィルタリングが出来ればそういったトラブルの抑制につながるのではないかと私は考えた。

2. 関連研究

参考にする関連した研究について記述する。

2.1　Web 上の誹謗中傷を表す文の自動検出

Web 上の多くの文章に対して誹謗中傷となる単語を自動検出するための研究が行われていた[1]。この研究では悪口となる単語を悪口極性の単語とし、またその逆極性となる全く関係ない単語等を用意し、その単語と判定したい対象の単語の Web 検索ヒット件数から悪口極性に強く共起しているかを調べることで悪口を検出しようとした。この手法は十分に有効であると報告している。

2.2　有害表現抽出に対する種単語の影響に関する一考察

Web 上の多くの文章に対して誹謗中傷等の有害表現を検出する手法に対して、既存の種単語を用いて Web 検索ヒット件数から有害極性となる単語を検出する手法について種単語の規模を拡張し、種単語の組み合わせ等が及ぼす影響についての検証が行われていた[2]。この研究では様々な方法や基準で種単語を選別し、その組み合わせによる種単語の有害極性判定手法の結果を示していた。結果として、種単語の組み合わせにより手法の平均精度は向上し、また、人手で判断した単語を種単語として用いると多くの有害語が高い極性値を持っており、精度に影響を及ぼすことが報告されている。

2.3　単語の分散表現を利用した文書類似度

ニューラルネットワーク言語モデルの一つである Word2Vec を用いた単語の分散表現をもとに文書を単語の集合として表現することで、文書の類似度を計算する研究が行われていた[3]。この研究では、Earth Mover’s Distance を用いて文書を単語で表わされるヒストグラムとみなすことで文書の類似度を求めていた。この手法によって同義語や類義語を考慮した文書間の類似度を求めることが出来ると報告している。

2.4　Twitter 上に投稿された文章に基づく感情推定法とその応用に関する検討

Twitter のツイートを対象に Word2Vec を用いてツイートのテキストを特徴ベクトルに変換し、ツイートから感情を推定する手法についての研究が行われていた[4]。この研究では喜、怒、哀、楽、無感情の５つの感情を表現する単語を用いてツイートの特徴ベクトルを生成し、ランダムフォレストを用いて分類を行っていた。結果としてテストデータではある程度の精度で分類できたが、一般的なツイートを対象とすると精度が落ちたと報告している

3. 提案する方法

3.1 Word2Vec を用いた悪口の検出手法について

既存の手法では、あらかじめ「キモイ」、「ダサ

い」といった悪口となる種単語と悪口とは関係ない種単語を用意し、入力単語と一緒に Web 検索をすることでそのヒット件数から悪口単語にどれだけ近いか、関連しているかを調べ、その関連度から悪口を検出する手法が提案されている[1]。しかし、Web 検索のヒット件数は不安定であり、再度実験を行った際、結果の精度が落ちているという研究報告も存在する[2]。その研究報告では種単語の選定によって改善するだろうとしていたが、種単語ではなく Web 検索のネットワークの方を改善することで精度の低下を防ぐことが出来ないかと考えた。

そこで、Word2Vec を用いて機械学習させたネットワークから単語間の類似度から、種単語との関連度を調べ、それを基に悪口を検出する手法を考えた。Web 上の文書は年々増加傾向にあり、従来のヒット件数を用いる手法では全く関係ない文書がヒット件数に影響する可能性がある。Word2Vec ではベクトル同士の演算によって似た意味を持つ単語を調べることが出来る[3]。また、将来言葉の意味や使い方が変化する、または新しい単語が生まれた場合、その単語について学習を行うことが出来るような文書集合が存在すればそれを基に学習させることで、言語の変化に対応できるのではないかと考えた。

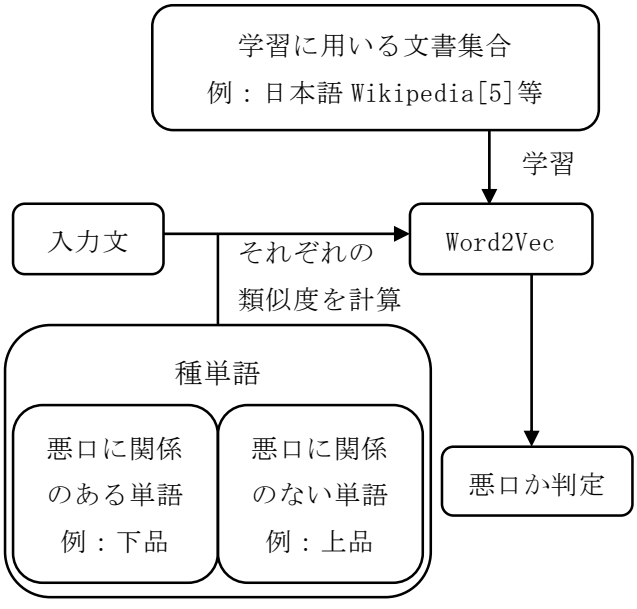

図 1.提案手法

3.2 悪口と感情分析の結果の関係の調査

Twitter のツイート文を対象にユーザの感情を推定する手法が存在する[4]。こちらの手法は Word2Vec を用いて喜、怒、哀、楽、無感情の５種類の感情に分類していた。私は、悪口と文に含まれている感情には何らかの関係があると考えた。よってこの手法を用いて文章の感情を推定し、悪口表現の検出手法の結果と合わせることで、感情分析の結果と悪口の関係についての調査することを考えた。

4. おわりに

今回は悪口表現の検出手法について、従来の手法であった悪口となる種単語との極性について、Word2Vec を用いて調べる手法を提案した。また、悪口表現とそれに伴う感情を推定することでその関係について調査を考えた。今後、この手法の実装を行い、テストを行う予定である。

参考文献

[1] 石坂達也，山本和英，“Web 上の誹謗中傷を表す文の自動検出，“言語処理学会第 17 会年次大会発表論文集，pp.131-134，2011.

[2] 畠山 鈴生，桝井 文人，プタシンスキ ミハウ，山本 和英，“有害表現抽出に対する種単語の影響に関する一考察，” 2016 年度人工知能学会全国大会（第 30 回），2016.

[3] 柳本 豪一，“単語の分散表現を利用した文書類似度，” 2015 年度人工知能学会全国大会（第 29 回），2015.

[4] 松林 圭，五味 京祐，古川 和祈，松尾 祐佳，松原 良和，日諸 マルセロ優次，中村 拓哉，山下 晃弘，松林 勝志，“Twitter 上に投稿された文章に基づく感情推定法とその応用に関する検討，” 第 78 回全国大会講演論文集，Vol1，pp.79-80，2016.

[5] 鈴木正敏，“日本語 Wikipedia エンティティベクトル，” http://www.cl.ecei.tohoku.ac.jp/~m-suzuki/jawiki_vector/，August 2019.