一般口演 | 第40回医療情報学連合大会（第21回日本医療情報学会学術大会） | 一般口演
一般口演18
自然言語処理
2020年11月21日(土) 11:15 〜 12:38 F会場 (イベントホール・特設会場1)

[4-F-1-02] 中国 SNS（ウェイボー）における新型コロナウイルス関連単語のテキスト分析とセンチメント分析

*曹 瀛丹[1]、張 洪健[1]、小笠原 克彦[1] （1. 北海道大学大学大学院保健科学院）
*YINGDAN CAO[1], Hongjian Zhang[1], Katsuhiko Ogasawara[1] （1. 北海道大学大学大学院保健科学院）
キーワード：COVID-19, Text mining, Sentiment analysis, China, Weibo

【背景】新型コロナウィルス感染症の流行に伴い、世界中の人々はこの感染症に対する不安を解消するために、SNS（ソーシャル・ネットワーキング・サービス）を利用して情報の収集と伝達や交流を活発に行っている。しかし， SNSの問題点として， SNS上には誤った情報が広まりやすくなるが、その情報伝達に関する研究は少ない。本研究では、中国を対象として、新型コロナウイルス感染症に関する SNS情報収集を通して、時系列的なテキスト分析とセンチメント分析を行い、市民が関心を有する関連単語頻度を統計と感情ポイントを算出した。【方法】中国 SNSでウェイボーを対象として、2020年1月から3月まで収集した。分析には Jiebaで中国語を形態素解析と TF-IDF法によって新型コロナウイルス関連単語頻度を統計解析し、ワードクラウドで新型コロナウイルス関連単語のワードクラウド図を作成した。 SnowNLPでひとつずつウェイボーをセンチメント分析をによる感情ポイント（最小値：0（ネガティブ感情）、最大値：1（ポジティブ感情））を算出し、平均値を算出した。【結果】本研究の結果として3月のみの結果を示す。一ヶ月の21742件ウェイボーが抽出された。関連単語のワードクラウド図から、単語の重要度と出現頻度を得られた。統計解析した頻度の比較において、新型コロナウイルス関連単語頻度のトップ10は「疫情（疫病流行）（出現頻度：5568）、新冠（新型コロナ）（4746）、肺炎（肺炎）（4591）、美国（米国）（3623）、防控（予防）（3035）、中国（2693）、工作（仕事）（2566）、确（診断）（2363）、病例（症例）（2274），开学（始業）（2124）」の順で高かった。感情ポイントの分布は、平均値が0.643（標準偏差：0.356）であった。

中国 SNS（ウェイボー）における新型コロナウイルス関連単語のテキスト分析とセンチメント分析

曹 瀛丹[*1]、張 洪健[*1]、小笠原 克彦[*1]
*1 北海道大学大学院保健科学院

Text Analysis and Sentiment Analysis of COVID-19 related words in Chinese Social Networking Services (Sina Weibo)

Yingdan Cao[*1] , Hongjian Zhang[*1] , Katsuhiko Ogasawara[*1]
*1 Graduate School of Health Sciences, Hokkaido University

Novel Coronavirus COVID-19 was first discovered in Wuhan City, Hubei Province in December 2019, and then spread rapidly to many countries around the world in early 2020. With the prevalence of COVID-19 infection, people all over the world are actively using SNS (Social Network Service) to collect, convey and exchange information in order to eliminate worries about this infection. However, as a problem point of SNS, although erroneous information on SNS is easy to spread, the research on this information transmission is very important. This study takes Chinese as the object, collects SNS information of infectious diseases in COVID-19, conducts time series text analysis and sentiment analysis, counts the relevant word frequency of public concern and calculates sentiment points. In the analysis, we used Jieba made a statistical analysis of the frequency of COVID-19-related words in Chinese through morpheme analysis and TF-IDF method. SnowNLP is used to analyze the sentiment of Weibo one by one, and the sentiment polarity (minimum value: 0 (negative sentiment), maximum value: 1 (positive sentiment)) are calculated. As the result of this study, 133,958 microblogs were extracted in a month of March. In the Term Frequency-Inverse Document Frequency of statistical analysis, the top 10 words in COVID-19 are " Confirmed, Cases, New, Virus, Italy, United States, Cumulative, Trump, Infection, Patients". Distribution of sentiment points, average value is 0.661 (standard deviation: 0.377).

Keywords: COVID-19, Text mining, Sentiment analysis, China, Weibo

1. 緒論

2019 年 12 月、中国湖北省武漢市で原因不明の肺炎の症例が報告され、1 月に新型コロナウイルス感染症(COVID-19)によるものと判明した。COVID-19 は新型コロナウイルスによって引き起こされる潜伏期の長い、伝染性の高い疾患である[1)]。2020 年 3 月 31 日 24 時、中国本土で報告された感染者は 81554 例、死者は 3312 例であった[2)]。2020 年 3 月には世界の多くの国に感染が広がった。

COVID-19 の流行に伴い、世界中の人々はこの感染症に対する不安を解消するために、SNS を利用して情報の収集、発信や交流を活発に行っている[3)4)]。一方、SNS の問題点として、誤った情報が広まりやすいことや、ネット世論の危機を招く可能性があることが挙げられるので、ネット世論を分析する必要がある[5)]。

ウェイボー（Sina Weibo）は、中国で最も人気があるソーシャルメディアプラットフォームの一つとして、ツイッターと同様に、ユーザーがリンク、画像、ビデオを追加してメッセージを送信または転送することができる。Weibo のアクティブユーザーは 5 億 5000 万人以上であり、COVID-19 の発生後、Weibo は「肺炎対策」議論エリアを開設し、毎日 2 億人以上のユーザーが Weibo で疫病の発生状況を追跡し、1 日平均 120 億回閲覧している[6)] 。

2. 目的

本研究の目的として、COVID-19 に関する Weibo 情報収集を通して、Weibo の内容をテキストと感情で分析することにより、市民の感情の変化や関心の内容を判断し、今後の政府や医療メディア等による情報の正確な伝達や市民の感情の変化の予測のための参考とする。

3. 方法

中国 SNS である Weibo を対象として、2020 年 3 月 COVID-19 に関する Weibo を収集した。中国語の形態素解析手法である Jieba と TF-IDF 法で COVID-19 関連単語の頻度や重要度を統計解析した。SnowNLP でそれぞれの投稿に対して感情分析をを行うために、感情極性値（最小値：0（ネガティブ感情）、最大値：1（ポジティブ感情））の平均値を算出した。

3. 1 データ収集

2020 年 3 月 1 日から 2020 年 3 月 31 日までの期間について、中国 SNS（Weibo）を用い、「新冠疫情（COVID-19 の発生状況）」と「新冠肺炎（COVID-19 による肺炎）」を検索キーワードとし、784,300 件の COVID-19 に関する Weibo を収集した。

収集した Weibo のうち、意味を持っていない（URL、画像、コメントなしの再投稿などの重複の内容）を除き、133,958 件の分析対象となる Weibo テキスト内容のノイズ（スペース、句読点、#タグ、@users など）を除去し、フィルタリングをした。

3. 2 テキスト分析

テキスト分析は Jieba で中国語の形態素解析を行い、TF-IDF 法によって COVID-19 関連単語の頻度を統計解析した。形態素解析とは、自然言語処理（Natural Language Processing）の一部で、自然言語で書かれた文を言語上で意味を持つ最小単位(＝形態素)に分け、それぞれの品詞や変化活用形などを判別することである。TF-IDF 法とは，Term Frequency(TF)と、Inverse Document Frequency(IDF)のことであり、単語に対する重み付けの指標の一種である。単語の出現頻度 TF と文書頻度の逆数 IDF の積により求める。

3.3 センチメント分析

センチメント分析とは、文字通り投稿者の「センチメント = 感情」を分析することを意味する。ウェブ上に投稿されたコメントなどを分析することによって、投稿者が持っている感情がネガティブなのかポジティブなのか、また、どの程度の強さなのかを知ることができる。

本研究ではセンチメント分析の手法である SnowNLP を用いた。SnowNLP での感情分析の結果は極性値のような形で返される。極性値のスコアは[0.0, 1.0]の範囲であり、1.0 がポジティブ感情、0.0 がネガティブ感情となる。

4 結果

本研究の結果として 3 月のみの結果を示す。COVID-19 関連単語の TF 値と TF-IDF 値の比較において、COVID-19 関連単語の出現頻度と TF-IDF 重要度トップ 10 を表 1 に示した。(検索キーワードを除いた。)TF 値による COVID-19 関連単語の出現頻度のトップ 10 は「确诊(診断)(出現頻度:82643)、病例(症例)(出現頻度:71709)、防控(予防)(出現頻度:59110)、中国(出現頻度:2693)、美国(米国)(出現頻度:43301)、新增(新規患者数)(出現頻度:35240)、工作(仕事)(出現頻度:34451)、意大利(イタリア)(出現頻度:30837)、累计(累積)(出現頻度:29031)、全球(世界)(出現頻度:26275)」の順で高かった。

表 1　3 月の COVID-19 関連単語トップ 10

	TF 値トップ 10 の単語	TF-IDF 値トップ 10 の単語
1	确诊(診断)	确诊(診断)
2	病例(症例)	病例(症例)
3	防控(予防)	新增(新規患者数)
4	中国(中国)	病毒(ウイルス)
5	美国(米国)	意大利(イタリア)
6	新增(新規患者数)	美国(米国)
7	工作(仕事)	累计(累積)
8	意大利(イタリア)	特朗普(トランプ)
9	累计(累積)	感染（感染）
10	全球(世界)	患者(患者)

3 月における COVID-19 関連の Weibo 内容の感情極性値の分布を図 1 に示す。

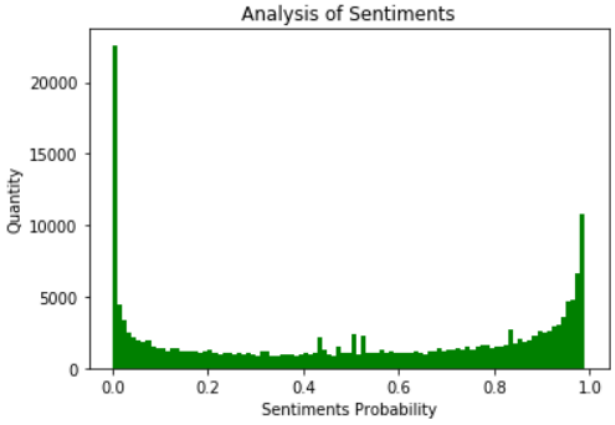

図 1　3 月 Weibo 内容の感情可能性分布図

3 月の感情極性値は、平均値が 0.661(標準偏差:0.377)であった。

5. 考察

本研究では Weibo の内容に対してテキスト内容と感情極性値分析を行い、市民が 3 月に関心があった内容や感情の変化を明らかにした。

TF 値と TF-IDF 値による重要度 TOP10 の単語から、人々がこの期間に最も関心を持っているのは、感染症の拡大状況であると考えられる。TF-IDF 値による重要度 TOP10 の単語の中には、「意大利(イタリア)、美国(米国)、特朗普(トランプ)」がある。それはこの時期に世界では COVID-19 が流行したため、人々が海外の流行状況に関心があったと考えられる。

分析期間であった 3 月は、COVID-19 に対する予防策などを多くの人々が理解していたため、COVID-19 に対する恐怖は治まっていたと思われる。一方，中国本土での症例数の増加速度は緩やかであり，中国での感染拡大時期は過ぎていたため，人々の態度はやや積極的な状態にあった。海外疫病の発生状況の動向も注目されているが、距離が離れているため、恐れや不安などのネガティブな感情はポジティブな感情に比べて少ないと考えられる。

本研究の限界として、本研究では Weibo のみからデータを収集した検討であったため、Weibo を利用していない人に対して、注目内容や感情を収集することができない。本研究では Weibo ユーザーの性別や年齢のデータを収集していないため、結果を年齢や性別による分析はできない。今後は感情分析とトピック抽出内容の分析を組み合わせて、人々の感情の変化をより深く分析していきたいと考える。

6. 結論

分析期間であった 3 月は、中国 COVID-19 に関する市民の関心の内容は、感染症の拡大状況や海外の流行状況を発見した。人々の関心の内容が 3 月に中国の疫病流行から世界各国の流行に移ったことが明らかになった。3 月には中国の流行終息時期ため、人々の態度はやや積極的な状態にあったことが明らかになった。

参考文献

1) Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. The Lancet 2020 Feb 15;395(10223):497-506 [FREE Full text] [CrossRef] [Medline]
2) National Health Commission of the people's Republic of China. Pneumonia situation in novel coronavirus infection on March 31st [in Chinese] [http://www.nhc.gov.cn/xcs/yqtb/202004/28668f987f3a4e58b1a2a75db60d8cf2.shtml(cited 2020-Agu-25)]
3) Zhao Y, Cheng S, Yu X, Xu H Chinese Public‘s Attention to the COVID-19 Epidemic on Social Media: Observational Descriptive Study J Med Internet Res 2020;22(5):e18825 [https://www.jmir.org/2020/5/e18825 DOI: 10.2196/18825 PMID: 32314976 PMCID: 7199804 (cited 2020-Agu-25)]
4) Chen, X., Lun, Y., Yan, J. et al. Discovering thematic change and evolution of utilizing social media for healthcare research. BMC Med Inform Decis Mak 19, 50 (2019). [https://doi.org/10.1186/s12911-019-0757-4(cited 2020-Agu-25)]
5) Househ M. Communicating Ebola through social media and electronic news media outlets: a cross-sectional study. Health Informatics J 2016 Sep;22(3):470-478. [CrossRef] [Medline]