Validating Sentiment Analysis on Opinion Mining Using Self-reported Attitude Scores

Jieyu Ding Featherstone
Department of Communication
University of California, Davis
Davis CA, USA
jding@ucdavis.edu

George A. Barnett
Department of Communication
University of California, Davis
Davis CA, USA
gabarnett@ucdavis.edu

2020 Seventh International Conference on Social Networks Analysis, Management and Security (SNAMS) | 978-0-7381-1180-3/20/$31.00 ©2020 IEEE | DOI: 10.1109/SNAMS52053.2020.9336540

***Abstract*—This paper aims to validate the use of sentiment analysis of text on opinion mining, using *IBM Watson's Natural Language Understanding*. An online survey examining attitude towards genomic editing using a validated scale and a text entry to comment on their thoughts of this technology was given to four populations—scientists, policymakers, farmers and the general public. The results indicate that sentiment scores of texts are significantly related to attitude scores in three of the four populations. This study shows that sentiment analysis is a reliable tool to understand opinions.**

Keywords—sentiment analysis, attitude, opinion mining, surveys, validation.

I. Introduction

Sentiment analysis, also known as opinion mining of texts, is a subdiscipline of natural language processing (NLP) and computational linguistics, refers to the techniques used to extract, classify, understand, and assess the opinions expressed in online texts [15]. Large amounts of online text data have facilitated the development of sentiment analysis especially in speculating people's attitudes, opinions, and beliefs [1].

Sentiment analysis has been widely used in various contexts to predict public opinions and trends. Scholars in political communication apply sentiment analysis on social media posts to extract public opinions on presidential candidates and accurately predict elections results [2]. Marketing researchers find that sentiment tracking on product reviews helps them understand the popularity of products and consumer preferences [3]. Studies in economics find that sentiment analysis of news reports predicts stock market trends [4].

Studies of sentiment analysis have highlighted the importance of opinion mining through texts in predicting human emotions, attitudes, and opinions. Online businesses use opinion mining of online reviews to extract insights of feedbacks on their products so that they can adjust their marketing strategies and improve their products and services [13]. Governments extract public sentiments from online discussions about certain policies and make policy changes according to public needs [14]. These studies show that sentiment of texts provide insights of public opinions which help decision makings.

Therefore, in order to accurately capture public opinions from sentiment analysis and improve the predictability of sentiment analysis on opinion mining, this paper examines the relationship between sentiment scores and attitude measures toward genomic editing from a survey conducted among four different populations, including scientists, policymakers, farmers and the general public. The survey included a validated 14-item attitude scale to examine general attitude towards genomic editing and a text entry question with around 100-word asking about thoughts on genomic editing. Then we used IBM Watson's *Natural Language Understanding* to analyze the sentiment scores of all the texts about genomic editing and conducted Pearson Correlation tests between sentiment and attitude scores.

The motivation behind the study comes from the lack of validation of sentiment analysis using an established measurement outside of machine learning tools. Past studies have been trying to validate the reliability of sentiment analysis using different machine learning tools and methods or validating it through human coding [1]. These studies have found ideal models and tools that generated accurate results of sentiments, however, they did not validate whether the detected results match the results of self-reported human attitudes. We believe that while it is important to validate and discover accurate machine learning tools for sentiment analysis, it is equally important to validate whether these sentiment scores accurately represent human attitude.

Therefore, this paper aims to validate sentiment analysis as a tool for opinion mining using self-reported attitude scores. This study will provide insights on how accurate sentiment analysis predicts opinions, which would help guide the use of sentiment analysis on texts. The paper is organized as follows: Section II explains the nature of sentiment analysis and methods of validation. Section III describes our study methods and results. Section IV is conclusion.

II. Sentiment Analysis

A. Sentiment Analysis and Human Attitudes

Sentiment analysis assumes that lexical items in the text carry attitudinal loadings. In other words, texts not only include factual expressions about entities, events, and their properties but also include opinion expressions that describe human sentiments, appraisals, or feelings toward entities, events, and their properties [5]. Therefore, the goal of sentiment analysis is to extract opinion expressions about human feelings towards these entities, events, and their properties from texts.

On the other hand, psychological research has extensively studied public opinions and terms it as human attitudes. Psychologists believe that attitude is evaluative and affective, meaning that attitude is what individuals judge whether they are for or against some object. Therefore, attitude can be measured by bipolar scales (i.e. object A is positive/good, neither good or bad, negative/bad) [6].

Funding: Innovative Genomics Institute Society Research Award

Since studies in psychology have found that attitude is an important predictor of decision making and sentiment analysis aims to extract human attitude through texts, this study links these two measures by examining the relationship between text extracted attitude and self-reported attitude (as defined in psychology). The correlation between the two can shed light upon the accuracy and utility of sentiment analysis of text in predicting human behaviors.

B. *Approaches of Sentiment Analysis*

There are two main approaches for automated sentiment analysis: unsupervised and supervised learning. Unsupervised learning uses a pre-defined dictionary of words, phrases, and sentences to classify whether a text is positive, negative, or neutral, the software (IBM Watson) used in this study is a type of unsupervised learning. Whereas supervised learning requires training sentiment classifier with texts that already have known classifications (i.e. rated online reviews that are negative, neutral, or positive). After training the machine, models are built to predict unlabeled texts [5]. This paper focuses on unsupervised learning as it is more cost-effective and widely used on large data sets.

Unsupervised learning, also known as lexicon-based learning, has been a popular method for automated content analysis. This method is straightforward in classifying documents as it calculates frequencies of words, their weightings in the documents, and aggregates the sentiment scores of words that are included in the dictionary [7]. The biggest challenge of sentiment analysis is the target of object. For example, the text "I love iPhone features and functions. My mom hates me using iPhone" has two distinct sentiments, one is "my" love for iPhone, which is positive and the other is "mom" hates it, which is negative. When calculating sentiment for the text, it is possible that the result is neutral since there are two opposite sentiments about iPhone. Therefore, in order to accurately capture the sentiment of our target, it is the best to tell machines the keywords we are analyzing around [5], which are: iPhone and I in the above example.

Therefore, this study uses IBM Watson's *Natural Language Understanding* (NLU) to analyze sentiment of texts. The NLU provides a step before sentiment analysis that allows researchers to detect different themes and features of each text and then focus on analyzing the sentiment of the specific feature (target). As a result, the sentiment scores are generally more accurate for the target [8].

C. *Validation of Sentiment Analysis*

Most sentiment analyses have been validated through two ways: (1) comparing the results with other algorithms; (2) comparing the results to human coding.

When comparing sentiment results either among different models/algorithms or to human coders, F1 score is an important measure used to examine the accuracy of classification. Researchers usually analyze sentiments of the same texts with different algorithms and compare F1 scores from different models with human coding. After comparisons, the model that performs the best (the highest F1 score) will be refined and attuned to the features and characteristics of the data set. For example, VADER was picked as the outstanding model and refined to adapt to the blog data set [9]. Usually, grammatical and syntactical features are adjusted to attune to the contexts in the model. These validation methods are great in understanding whether machines understand the texts as the way humans do. However, these methods do not tell whether the sentiment of the texts represents or to what extent it represents self-reported attitude towards a certain object we are measuring. As this paper argues self-reported attitudes are widely studied as an important predictor of human behaviors. Therefore, this study validates sentiment analysis results with a self-reported attitude measure.

III. STUDY METHODS AND RESULTS

A. *Methods*

A survey about attitude towards genomic editing was collected among four populations in 2018: farmers, scientists, policy makers, and the general public. Both the general public and farmers samples were recruited through Qualtrics panels (www.qualtrics.com), an online sampling and survey platform [10]. Qualtrics collected representative samples of the U.S. scientists (genetics and genomics faculty at major U.S. land grant universities) and policy makers (staff at federal- and state-level agricultural policymaking institutions) were recruited from two sampling frames. Participants were asked to fill out their attitude towards gene editing on a validated scale [11]. Also, they were given definitions of the terms "genome editing" and "genome" and asked to compose a short essay (100+ words) on their thoughts and opinions on genome editing and CRISPR.

The definition prompt states: "Genome: The term "genome" encompasses all of an organism's genetic material, or DNA. It can be thought of as the instruction manual for living things, including plants, bacteria, and animals. Genome editing: Genome editing describes a range of techniques that make it possible to alter a selected part of the genome in a living cell by removing or changing existing elements or adding new ones to changes in physical traits and prevent disease. Scientists use different technologies to do this. These technologies "cut and paste" the DNA at a specific spot, allowing scientists to remove, add, or replace the DNA. Recently, a new genome editing tool called CRISPR, has been developed. Many scientists who perform genome editing now use CRISPR."

The reliability of the attitude scale was calculated using principal components analysis (PCA) in SPSS 26V. The PCA indicated the number of factors and the score of each item under these factors. Factors represented the number of aspects measured under attitude and the score of each item indicated how well the item represented this factor. Items with scores under 0.75 were deleted from the scale. The final scale had 16 reliable items to indicate attitude towards gene editing. Then the scale was aggregated by taking the average of the 16 items.

Lastly, we cleaned the texts by removing stop words and unique symbols [12] and we analyzed sentiments of texts using Watson's API service in Python.

After we have both sentiment scores and attitude scores, we conducted Pearson's correlation between the two scores separately among the four populations.

B. Results

Reliability tests showed that the attitude scale reached an average of $\alpha = .93$ among all four populations. TABLE 1 shows the attitude scores for each population, the higher the score, the more positive the attitude. TABLE 1 reveals that attitude scores are higher among scientists and policy makers whereas lower among general public and farmers. Moreover, the variability of attitude scores is much lower among scientists than the other three populations. TABLE 1 also shows the reliabilities for the individual groups.

TABLE I. AVERAGES AND STANDARD DEVIATIONS OF ATTITUDES TOWARDS GENE EDITING AMONG FOUR POPULATIONS

Sample	Attitude towards gene editing			
	Sample size	***Mean***	***St. Deviation***	***α***
Scientist	153	4.59	.37	.85
General public	485	3.89	.70	.91
Farmer	168	3.74	.77	.95
Policy maker	71	4.32	.67	.93

TABLE 2 shows the percentages, average scores and standard deviations of sentiment scores. Sentiment scores range from -1 to 1, a score from -1 to 0 indicates negative sentiment, from 0 to 1 indicates positive sentiment, and 0 indicates neutral sentiment. Note that scientists had the most positive (and least negative) sentiment, followed by the policy-makers. Farmers were more neutral, less positive and negative than the general public who were bipolar in their sentiment being both more negative than the other groups but slightly more positive than the farmers.

TABLE II. PERCENTAGES OF SENTIMENT LABELS, AVERAGES AND STANDARD DEVIATIONS OF SENTIMENT SCORES

Sample	Sentiment labels and scores				
	Negative	***Neutral***	***Positive***	***Mean***	***St. Deviation***
Scientist	9.2%	3.9%	86.9%	.54	.40
General public	32%	5.4%	62.6%	.24	.62
Farmer	28.6%	11.9%	59.5%	.25	.62
Policy maker	19.7%	9.9%	70.4%	.36	.54

TABLE 3 shows Pearson's correlation (2-tailed) between attitude scores and sentiment scores. The higher the score, the more correlated attitudes and sentiments are, indicating the sentiment analysis is more reliable in detecting human attitude.

TABLE III. PEARSON CORRELATION BETWEEN ATTITUDE SCORES AND SENTIMENT SCORES

Sample	Pearson Correlation (*r*)	
	r	**Significance (*p*<.05)**
Scientist	.01	.913
General public	.31	.000
Farmer	.38	.000
Policy maker	.23	.05

The results show that except for scientist sample, sentiment scores have a significant correlation with attitude scores, indicating that sentiment of texts about gene editing has a small to medium predicting power to extract attitude toward gene editing among the general public, farmer, and policy maker samples. We believe that the low variability of attitude and sentiment scores among scientist sample makes it hard to detect any correlation as there are not sufficient variance on the attitude scale or the sentiment measure.

IV. CONCLUSION

This paper uses self-reported attitude scores to validate the power of sentiment analysis for opinion mining. Results show that sentiment analysis has a small to medium correlation with self-reported attitude scores among general public, farmers, and policy makers. This shows sentiment analysis has some prediction power to indicate human attitudes, however, less than 15% of the variance in attitude could be accounted for by Watson's sentiment analysis. The low effect size implies that sentiment analysis using IBM Watson's NLU has limitations on the extent to which it explains attitudes. This result shows the significance of using sentiment analysis as an opinion mining tool while reminding users of the limited extent to which sentiment analysis indicates public opinions. Future research should (1) refine sentiment analysis tools so that they can also accurately signal public opinions and (2) discover other methods to measure public opinions.

The correlation between the attitude measure and sentiment analysis scores was not significant for the scientist sample. It is possible that the low variability of attitude and sentiment scores among scientists makes it difficult to detect any correlation as there are not sufficient variance on either the attitude scale or the sentiment measure. This result indicates the importance of sampling even when analyzing big data. Future research should explore how different samples could influence the predictive power of sentiment analysis.

Additional research on the same samples found that scientific knowledge was a significant predictor of attitudes toward gene editing [16]. This suggests that there may be other factors in addition to attitudes that may determine sentiment as measured by Watson. Future research should also determine how other variables influence sentiment.

This paper is the first to explore the relationship between sentiment analysis scores of texts and their corresponding self-reported attitude scores on the same subject. This research bridges the attitude concept in psychology with the state-of-art sentiment analysis on texts. Our study implies that using sentiment analysis to extract public opinions is promising and requires delicate methods for validation.

REFERENCES

[1] S. Gonzale-Bailon, and G. Paltoglou, “Signals of public opinion in online communication: A comparison of methods and data sources,” Ann Am Acad Polit Ss, vol. 659, pp. 95-107, May 2015.

[2] F. Nausheen, and S. H. Begum, “Sentiment Analysis to Predict Election Results Using Python,” *Proceedings of the 2nd International Conference on Inventive Systems and Control (Icisc 2018)*, pp. 1259-1262, 2018.

[3] S. Ahmed, and A. Danti, “Effective Sentimental Analysis and Opinion Mining of Web Reviews Using Rule Based Classifiers,” *Computational Intelligence in Data Mining, Vol 1, Cidm 2015,* vol. 410, pp. 171-179, 2016.

[4] R. Ren, D. D. Wu, and T. X. Liu, “Forecasting Stock Market Movement Direction Using Sentiment Analysis and Support Vector Machine,” *Ieee Systems Journal,* vol. 13, no. 1, pp. 760-770, Mar, 2019.

[5] B. Liu, *Sentiment analysis and opinion mining*, Chicago, IL: Morgan & Claypool, 2012.

[6] M. Fishbein, and I. Ajzen, *Belief, Attitude, Intention, and Behavior*, Reading, MA: Addison-Wesley Publishing Company, 1975.

[7] B. Pang, and L. Lee, “Opinion mining and sentiment analysis,” *Foundations and Trends in Information Retrieval,* vol. 2, no. 1-2, pp. 1-135, 2008.

[8] IBM, *The Era of Cognitive Systems: An Inside Look at IBM Watson and How it Works*, Online: Redbooks, 2012.

[9] C. J. Hutto, and E. Gilbert, "VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text."

[10] C. J. Calabrese, J. D. Featherstone, M. Robbins, and G. A. Barnett, “Examining the Relationship Between Gene Editing Knowledge, Value Predispositions, and General Science Attitudes among U.S. Farmers, Scientists, Policymakers, and the General Public,” *Journal of Science Communication,* vol. submitted for publication, 2020.

[11] C. Funk, and M. Hefferon, *Public views of gene editing for babies depend on how it would be used*, Pew Research Center, 2018.

[12] J. Diesner, “ConText: Software for the Integrated Analysis of Text Data and Network Data,” in Preconference at Conference of International Communication Association (ICA), Seattle, WA, 2014.

[13] M. Hu, and B. Liu, “Mining and summarizing customer reviews,” in *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2004.

[14] G. Angulakshmi, and R. Manicka Chezian, “An analysis on opinion mining: techniques and tools,” *International Journal of Advanced Research in Computer and Communication Engineering*, vol. 3, no. 7, July 2014.

[15] H. Chen, and D. Zimbra, “AI and opinion mining,” *IEEE Intelligent Systems*, vol. 25, no. 3, May 2010.

[16] C. J. Calabrese, J. D. Featherstone, M. Robbins, and G. A. Barnett, “Examining the relationship between gene editing knowledge, value predispostions, and general science attitudes among U.S. farmers, scientists, policymakers, and the general public,” *Journal of Science Communication,* in press.