International Journal of Business, Human and Social Sciences
ISSN: 2517-9411
Vol:13, No:1, 2019

Assessment of the Validity of Sentiment Analysis as a Tool to Analyze the Emotional Content of Text

Trisha Malhotra

***Abstract*—**Sentiment analysis is a recent field of study that computationally assesses the emotional nature of a body of text. To assess its test-validity, sentiment analysis was carried out on the emotional corpus of text from a personal 15-day mood diary. Self-reported mood scores varied more or less accurately with daily mood evaluation score given by the software. On further assessment, it was found that while sentiment analysis was good at assessing 'global' mood, it was not able to 'locally' identify and differentially score synonyms of various emotional words. It is further critiqued for treating the intensity of an emotion as universal across cultures. Finally, the software is shown not to account for emotional complexity in sentences by treating emotions as strictly positive or negative. Hence, it is posited that a better output could be two (positive and negative) affect scores for the same body of text.

***Keywords*—**Analysis, data, diary, emotions, mood, sentiment.

I. Introduction

SENTIMENT analysis –or opinion mining– is the field of study that computationally analyzes people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities (products, services, organizations, individuals, issues, events, topics) and their attributes [4]. Statistical validity is an assessment of how 'well-founded' a measurement tool is and likely accurately corresponds with the real world based on probability [2]. Test validity is the degree to which preexisting theory and evidence support the interpretations of test scores [2]. In this report, the principles of validity and test validity are used to assess the usefulness of sentiment analysis as a measurement tool. To do so, sentiment analysis was carried out on the text from a personal 15-day 'mood diary' with a self-reported mood score (from 1-7) at the end of each entry.

Sentiment analysis package (available online) was loaded using R. Results of the experiment are as follows. On correlating self-reported mood scores with sentiment analysis mood scores (R software score), a strong positive correlation was seen, $r\ (13) = 0.8333$, $p < 0.001$, 95% CI [0.56, 0.94]. This means that the R-score linearly increased with an increase in the self-reported mood score and vice versa, as shown in Fig. 1. This occurred less than 5% of the time due of random factors or chance. While there could be other variables that moderate or mediate the behavior of either of these variables, we can conclude that the sentiment analysis more or less accurately varied with the self-reported mood scores for 15 entries.

Trisha S. Malhotra is a third year undergraduate student in the Department of Psychology at Ashoka University in Sonepat, Haryana (phone: 9022748948; e-mail: trisha.malhotra_ug19@ashoka.edu.in).

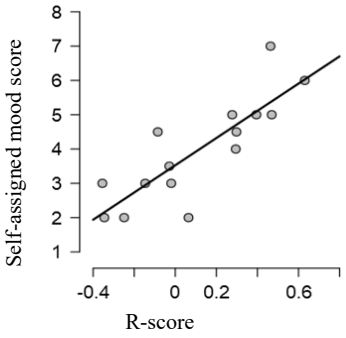

Fig. 1 Graph depicting strong positive correlation between self-reported mood scores (Y-axis) and R scores (X-axis)

II. Hypothesis

The researcher wanted to assess how 'well-founded' and therefore valid the overall function driving sentiment analysis was. Two popular functions involved are 'sentiment polarity', which calibrates the number of negative or positive words used [3], and the 'part of speech' (POS) function which looks for the nine parts of speech in a sentence. They are the noun, pronoun, adjective, conjunction, determiner, verb, adverb, prepositions and interjections. There is evidence suggesting that the POS function was not effective in the micro-blogging domain due to the informal, grammatically incorrect nature of the text [5]. Results from the study indicated that although the overall score was successfully impacted by the number of words (polarity), the software was not able to successfully identify and calibrate scores using the other important aspects of a sentence (POS function) when assessing free-form text from websites like Twitter.

In light of these findings, the researcher predicted that software would fail to identify emotive words used in the opposite context- 'incongruent sentences' in her own diary entries. To test this, the number of negative and positive words per entry was tabulated irrespective of the context in which they were used. 'Negative', as can be seen in Appendix Table V.A, refers to emotionally charged words having a negative affect (e.g.: frustrated, irritated, bad, anxious, crying, sad, etc.). Alternatively, 'positive', as seen in Table V.B, refers to emotionally charged words with positive affect (e.g.: happy, excited, productive, accomplished, fulfilled, calm, satisfied, etc.), while Table VI.A shows the number of negative and positive words per entry. 'Irrespective of context' implies that even if charged words were used in sentences that expressed the opposite intention– "I was not irritated" or "I was not very happy" – they were counted as correct uses of the word in an

entry. Hence, the researcher hypothesized that the number of negative words per entry (including those within incongruent contexts) would predict:

(i) a significantly lower R-mood score, and
(ii) no significant effect on the self-reported mood score.

III. Results

The following statistical tests were carried out using JASP. The average number of context-word incongruent sentences for negative words ("I am no longer sad; I feel kind-of sleepy instead.") was about 1 (1.061) sentence per entry, SD = 1.387, while the average number of sentences in an entry was 7 (6.773), SD= 1.880. A simple linear regression was used to predict the R-score of an entry based on the number of negative words used in it.

A significant regression equation was found, $F(1, 13) = 36.32$, $p < 0.001$ with an R^2 of 0.736. This means the number of negative words used did affect the R-score to a significant degree. To elaborate, the R-mood score decreased by 0.091 for each negatively charged word used in the entry. The result of the number of positive words on R-score was also significant with R's score rising by 0.090 for each positive word used (see the Appendix for Linear Regression Tables I and II).

IV. Discussion

This finding verifies the first part of the hypothesis. One could assume that R's polarity function accurately assessed the number of words but its POS function did not factor in their context, especially since a lot of the diary was in free-form. However, this was disputed when significant results were also observed for a regression between positively or negatively charged words and self-reported mood scores. Self-reported scores increased or decreased significantly depending on how many positive or negative words were used respectively. So, the first part of the hypothesis may be accepted but the second part is rejected. See the Appendix for linear regression Tables III.B and IV.B. Instead, either of two possibilities is revealed. The first is that R's POS function might have accounted for the incongruent sentences; however, since such sentences did not comprise of a significant portion of the entry, it did not affect the overall score. The second possibility is that even though many of the affective words might have been used in an opposing context; they nevertheless 'locally' played a significant role in impacting the researcher's intuitive guess at their 'global' mood. Hence, one cannot critique sentiment analysis for assigning valence to certain incongruent sentences instead of a neutral (0) score if one carries out that behavior.

Each of these possibilities inevitably supports the validity – the extent to which it accurately corresponds with the real world– of sentiment analysis. The findings suggest that its functions successfully imitate our perception of our own mood. However, this paradigm is not without its shortcomings.

V. Critique

A. Cross-Cultural Differences in Reported Sentiment Polarity

Firstly, the inability to incorporate pre-existing models about emotion –low test validity– serves as a critique of sentiment analysis. Sentiment analysis factors-in the difference between the "I am very x" and "I am x", where 'x' is a particular emotion. It does so by accounting for differences between phrases like "I am very happy" (average sentiment = 0.675) which is rightly interpreted as more positive than "I am happy" (average sentiment = 0.375). This is part of the aforementioned 'sentiment polarity'. However, the function does not deduce differences in subtleties between the affective intensity of the variations (synonyms) of a particular emotional word (when 'x' is replaced with a close cousin 'y'). For example: the sentence 'I am angry' yields an average sentiment score of -0.433127. However, if 'angry' is replaced with 'infuriated', 'enraged' or 'frustrated', all of which differ significantly [7], the sentiment polarity score remains the same. In general, sentiment analysis pragmatically ascribes a single, fixed score to every emotion. It relies on the theoretical assumption that certain emotions are basic and experienced universally across cultures.

Cross-cultural research brings into question the strict positivity and negativity of presumed universal emotions because an emotion that is considered negative in one culture can be positive in another culture. In an experiment by An et al., perceived emotional affect to so-called 'universal' emotions varied significantly cross-culturally [1]. More specifically, their results suggested that basic emotions like happiness or sadness contain levels of both positivity and negativity and there is a significant variation in the intensity of those levels across cultures. This affect-mismatch was observed when Koreans and Chinese participants reported stronger positivity to sadness compared to Canadian and American participants, while the latter reported stronger negativity of sadness compared to the former. Findings replicating cultural affect-mismatch have been observed in other studies as well [9].

The rationale for this difference in cultural perception is highlighted using the Component Processing Model (CPM) of emotions which claims that emotions have both affective (automatic) and cognitive (controlled) components [8]. It follows that affective components are more emotionally polarized than cognitive components. So for a positive emotion, a focus on the affective component would yield a more positively perceived emotion as opposed to a focus on the cognitive component. The difference between cultures has to do with which component they focus on when they perceive a particular emotion. Hence, some cultures may perceive anger and sadness as more positive because of an implicit focus on its cognitive component like Easterners did in An et al. [1]. This is in comparison to Westerners who perceived anger as more negative because of a focusing on its affective component. In light of these findings, it is reasonable to conclude that sentiment analysis does not account for the

cultural relativity in affect perception since it ascribed fixed values to each emotion. It has not incorporated CPM as a model, thereby reducing its test-validity.

B. Inability to Analyze Dialectical Text

In line with this critique, a criticism of sentiment analysis is that it lacks the incorporation of both a positive as well as negative affect score to a single corpus of text. Instead, it yields a single conglomerate score. The benefit of two scores would be that a more precise interpretation of opinions would be made possible. This is because there are situations in which people report feeling opposite emotions simultaneously. For instance, when adolescents leave their home for college they report feeling both happy and sad. This experience can occur more often in some cultures known to have a 'dialectical emotional style'–defined as "the propensity to experience both positive and negative emotion over time" [6]. For instance, it was found that the dialectical emotional experience is more prevalent in East Asians than Westerners who report feeling emotions more non-dialectically [6].

The problem with the sentiment analysis of a dialectical body of text (for opposite emotions) would be that the overall score would always regress towards 0 since only a single score is given as output. Such a score indicates that the positive feelings are dampened by negative ones. However, this 'dampened feeling' does not occur in the subjective experience of feeling two emotions at the same time. Both emotions are experienced just as powerfully and it is debatable whether one necessarily mitigates the impact of the other. A score of 0 implies the lack of any feeling (apathy), which is not the case. The interpretation of such a score could be incorrect if dialecticism is not taken into consideration. A possible improvement, therefore, would be the incorporations of a negative and positive score for the same corpus of text.

C. No Rationale for Results

Finally, a more general critique of sentiment analysis is regarding the nature of its approach. While it can help assess the opinions, attitudes and appraisals of a general population, it cannot provide a rationale for its findings. An interview or survey, although more time-consuming, can generate possible reasons behind the attitudes and opinions of users towards a particular product, event, or person. Then, by translating it into a question and collecting subjective responses, the survey can generalize the reasons behind a particular reaction of the population. Such a rationale would provide producers with integral feedback on what changes or improvements need to be made to their product. It can also explain why a particular event occurred. Sentiment analysis only provides a summary of people's feelings but cannot provide the primary reason behind them.

VI. Conclusion

Sentiment analysis is valid to the extent that through the researcher's own tests, it was found to successfully imitate and therefore capture how people perceive their own emotions. However, its (test) validity as a measurement tool is questioned as it fails to incorporate the Component Processing Model of emotion by assigning fixed scores to every emotion. CPM is the underlying mechanism through which cultures experience relativity in how they perceive what were believed to be universal emotions.

Sentiment analysis also only tells us half the story since it provides the polarity of opinion without providing any reasons behind its findings. This challenge can be tackled by conducting sentiment analysis in conjunction with large-scale surveys. Finally, a conglomerate score, as its output does not address the problem of input that is 'emotionally complex' or with two or more opposing emotions. A possible improvement to the design could, therefore, be to make sentiment analysis yield two (both negative and positive) scores for the same body of text which would successfully account for dialecticism of emotion.

Appendix

TABLE I.A
JASP Linear Regression Table Predicting R Score from Number of Positive Words Used

Model Summary				
Model	R	R^2	Adjusted R^2	RMSE
1	0.800	0.640	**0.612**	0.197

ANOVA

Model		Sum of Squares	df	Mean Square	F	p
1	Regression	0.895	**1**	**0.895**	**23.11**	**<0.001**
	Residual	0.503	13	0.039		
	Total	1.398	14			

TABLE I.B
Coefficients of ANOVA Predicting R Score from Number of Positive Words Used

Model		Unstandardized	Standard Error	Standardized	t	p
1	(Intercept)	-0.218	0.085		-2.556	0.024
	positive words	**0.090**	0.019	0.800	4.807	<0.001

TABLE II.A
JASP LINEAR REGRESSION TABLES PREDICTING R-SCORE FROM NUMBER OF NEGATIVE WORDS USED

Model Summary				
Model	R	Adjusted R^2	RMSE	
1	0.858	0.736	**0.716**	0.168

ANOVA						
Model		Sum of Squares	df	Mean Square	F	p
1	Regression	1.030	**1**	**1.030**	**36.32**	**<0.001**
	Residual	0.369	13	0.028		
	Total	1.398	14			

TABLE II.B
COEFFICIENTS OF ANOVA PREDICTING R SCORE FROM NUMBER OF NEGATIVE WORDS USED

Model		Unstandardized	Standard Error	Standardized	t	p
1	(Intercept)	0.419	0.067		6.243	<0.001
	negative words	**-0.091**	0.015	-0.858	-6.026	<0.001

TABLE III.A
JASP LINEAR REGRESSION TABLES PREDICTING SELF-ASSIGNED MOOD SCORE FROM NUMBER OF NEGATIVE WORDS USED

Model Summary				
Model	R	R^2	Adjusted R^2	RMSE
1	0.741	0.550	**0.515**	1.049

ANOVA						
Model		Sum of Squares	df	Mean Square	F	p
1	Regression	17.44	**1**	**17.438**	**15.86**	**0.002**
	Residual	14.30	13	1.100		
	Total	31.73	14			

TABLE III.B
COEFFICIENTS OF ANOVA PREDICTING SELF-ASSIGNED MOOD SCORE FROM NUMBER OF NEGATIVE WORDS USED

Model		Unstandardized	Standard Error	Standardized	t	p
1	(Intercept)	5.234	0.418		12.528	< .001
	negative words	**-0.373**	0.094	-0.741	-3.982	0.002

TABLE IV.A
JASP LINEAR REGRESSION TABLES PREDICTING SELF-ASSIGNED MOOD SCORES FROM NUMBER OF POSITIVE WORDS USED

Model Summary				
Model	R	R^2	Adjusted R^2	RMSE
1	0.838	0.703	**0.680**	0.852

ANOVA						
Model		Sum of Squares	df	Mean Square	F	p
1	Regression	22.306	**1**	**22.306**	**30.76**	**<0.001**
	Residual	9.428	13	0.725		
	Total	31.733	14			

TABLE IV.B
COEFFICIENTS OF ANOVA PREDICTING SELF-ASSIGNED MOOD SCORE FROM NUMBER OF POSITIVE WORDS USED

Model		Unstandardized	Standard Error	Standardized	t	p
1	(Intercept)	2.325	0.369		6.308	<0.001
	positive words	**0.448**	0.081	0.838	5.546	<0.001

TABLE V.A
LIST OF NEGATIVE AFFECTIVE WORDS USED IN ANALYSIS REPEATEDLY COUNTED AS USED ACROSS ENTRIES

1	Rough	18	Exasperated
2	Frustrated	19	Insecure
3	Anxious	20	Lazy
4	Irritated	21	Worthless
5	Stressed	22	Sad
6	Vigilant	23	Lonely
7	Annoyed	24	Disappointed
8	Frustrating	25	Distracted
9	Anger	26	Tired
10	Angst	27	Worse
11	Confusion	28	Scary
12	Worry	29	Nervous
13	Attacked	30	Afraid
14	Hurt	31	Weak
15	Crying	32	Demands
16	Bothering	33	Bad
17	Low		

TABLE V.B
LIST OF POSITIVE AFFECTIVE WORDS USED IN ANALYSIS REPEATEDLY COUNTED AS USED ACROSS ENTRIES

1	Accepted	18	Fun
2	Relief	19	Lovely
3	Calm	20	Grateful
4	Happy	21	Good
5	Enjoyable	22	Easy-going
6	Focused	23	Comfortable
7	Productive	24	Better
8	Accomplished	25	Helpful
9	Improve	26	Satisfied
10	Pretty	27	Hopeful
11	Optimistic	28	At-ease
12	Excited	29	Joyful
13	Relax	30	Fulfilling
14	Calmly	31	Pleasant
15	Confidence		
16	Positive		
17	Amazingly		

TABLE VI
TABULATED SELF-ASSIGNED MOOD SCORE, R-SCORE, WORD COUNT, SENTENCE COUNT, NEGATIVE AND POSITIVE WORDS USED FOR EACH ENTRY (15 SUCH ENTRIES)

Self-score 0-7	R-score	Word count	Sentence count	'Negative' words used	'Incongruent' negative sentences
3	-0.0195	109	7	6	1
6	0.63068	55	5	0	0
4.5	-0.085	129	9	4	1
2	0.06464	103	7	5	2
2	-0.3455	73	6	6	1
5	0.27732	97	9	1	0
7	0.46373	64	7	0	0
5	0.46973	73	6	1	0
3	-0.1463	75	8	7	4
3.5	-0.0288	84	9	4	2
4.5	0.29778	67	8	1	0
5	0.39489	59	5	0	0
4	0.29452	71	6	3	0
2	-0.2489	22	2	3	1
3	-0.355	92	7	10	4

REFERENCES

[1] An, S., Ji, L., Marks, M., & Zhang, Z. (2017). Two Sides of Emotion: Exploring Positivity and Negativity in Six Basic Emotions across Cultures. *Frontiers in Psychology,8*. doi:10.3389/fpsyg.2017.00610.

[2] American Educational Research Association, American Psychological Association &National Council on Measurement in Education. (1999). Statistics for educational and psychological testing. Washington DC: American Educational Research Association.

[3] B. (2016, June 16). Differences between Polarity and Topic-based Sentiment Analysis. https://blog.bitext.com/polarity-topic-sentiment-analysis.

[4] Bing Liu. (2012, May) Sentiment Analysis and Opinion Mining, Morgan & Claypool Publishers.

[5] Kouloumpis, Efthymios& Wilson, Theresa & Moore, Johanna. (2011). Twitter Sentiment Analysis: The Good the Bad and the OMG! ICWSM.

[6] Miyamoto, Yuri & D Ryff, Carol. (2011). Cultural differences in the dialectical and non-dialectical emotional styles and their implications for health. Cognition &emotion. 25. 22-39. 10.1080/02699931003612114

[7] S, P. (2011, January 08). Difference Between Anger and Rage. http://www.differencebetween.net/miscellaneous/difference-between-anger-and-rage/.

[8] Scherer K. R. (1982). Emotion as a process: function, origin, and regulation. *Soc. Sci. Inform.* 21555–570. 10.1177/053901882021004004.

[9] SimsT., Tsai J. L., Jiang D., Wang Y., Fung H. H., Zhang X. (2015). Wanting to maximize the positive and minimize the negative: implications for mixed affective experience in American and Chinese contexts. *J. Pers. Soc. Psychol.* 109 292–315. 10.1037/a0039276.