Original scientific paper / Izvorni znanstveni rad
Manuscript received: 2017-06-13
Accepted: 2017-08-22
Pages: 1 - 12

Textual Content and Engagement Correlation Analysis with Naive Bayes

Tomislav Krištof
Visoko učilište Algebra
Zagreb, Croatia
tomislav.kristof@algebra.hr

Vanja Šebek
Visoko učilište Algebra
Zagreb, Croatia
vanja.sebek@algebra.hr

Mario Fraculj
HURA
Zagreb, Croatia
mario.fraculj@hura.hr

Abstract: With the constant improvement of sentiment analysis software, it is possible to determine whether there is a correlation between the sentiment of the content and the content engagement. By combining two platforms we were able to prove that there is a moderate correlation between the content sentiment and content engagement. Furthermore, there are other correlations regarding numeric variables describing the properties of the content, like content length and title length compared to the content consummation and engagement. Determined values are showing strong negative correlation between the content length and content consummation. Content platform was Medium.com social network [22] and software platform for sentiment determination was an online tool [1] based on enhanced Naïve Bayes model [2]. For finding correlations we used the Pearson's correlation coefficient because it gives information about the magnitude of the association, or correlation, as well as the direction of the relationship.

Keywords: natural language processing, sentiment analysis, Pearson's correlation coefficient, enhanced Naïve Bayes model

INTRODUCTION

Sentiment analysis aims to measure positivity or negativity of written or spoken language. The positivity or negativity of analyzed text is determined through extraction of subjective information. The simplified model determines only polarity, while there are models that beside the polarity determine intensity [3].

Usage of sentiment analysis through NLP, natural language processing, is very popular and necessary in the realm of marketing. The engaged parties tend to determine sentiment and thus conclude whether the brand or the product perception is positive or negative. By measuring sentiment polarity of the brand through time it is possible to measure the effectiveness of marketing efforts, as well as quantify ROI.

The sentiment analysis software first has to learn to quantify sentiment. To do that it has to analyze data that possesses information about the sentiment. For the purpose of training the software the available data source was used [4] [5]. It is a dataset of highly polar movie reviews published on International Movie Data Base [6].

For the purpose of this work we have analyzed posts on popular social network Medium.com in order to discover whether there is correlation between the sentiment and post engagement. If the post sentiment was positive, would the readers tend to recommend it to other readers within their network?

Furthermore, how would the readers react if the post title is positive or negative or would they skip the post if it was too long.

Necessary steps:

- Provide a trustful sentiment analysis software [1].
- Provide sufficient volume and quality of written text [7] that can be quantified and measured by the means of qualitative descriptors (sentiment) and quantities (number of views, number of reads, read ratio, number of recommends, recommend-read ratio). The texts are compiled and suited to the specific audience in tags of “mental health” and “love” for which the author has top references. The author, Tomislav Krištof, has the *Top writer* status for these tags, which puts him among the top 50 writers for both categories in the world [9][10].
- Provide a tool to determine correlation (Microsoft Excel, Pearson’s correlation coefficient)
- Provide a tool for numeric and graphic presentation (Microsoft Excel)

Natural language processing [8] is an area of research and application that explores how computers can be used to understand and manipulate natural language text or speech to do useful things [11].

A number of researchers have attempted to come up with improved technology for performing various activities that form important parts of NLP works [11]. These works may be categorized as follows:

- Lexical and morphological analysis, noun phrase generation, word segmentation, etc. [12].

- Semantic and discourse analysis, word meaning and knowledge representation [13].
- Knowledge-based approaches and tools for NLP [14].

Sentiment analysis of language (written or spoken) uses three approaches [15]:

- knowledge-based techniques
- statistical methods
- hybrid approaches.

The aim of sentiment analysis is to determine polarity of the text by identifying words that describe emotional state. These states can be positive (happy, joyful), or negative (angry, sad). By diving into the context under which these identifiers are used, a learned processor can determine overall sentiment with great accuracy. Knowledge-based techniques determine the polarity of text by categories of affection founded on emotional words such as thrilled, depressed, scared [16].

SENTIMENT ANALYSIS TOOL USED

In order to produce reliable result in sentiment analysis it is necessary to have a well-trained software. To train the software one has to have a dataset that is of sufficient volume and polarity that can be determined with ease. For the learning of the tool [1], an available dataset [4] was used which was compiled by Andrew Maas et al. It is a set of 25000 highly polar text, movie reviews that include wide and various emotion words. The mentioned set was used for training. The other set of further 25000 words was used for testing the tool. All texts originate from International Movie Data Base [6]. Furthermore, most existing research on sentiment classification uses movie review data for benchmarking. The analysis tool used the 25,000 documents in the training set to build our supervised learning model. The other 25,000 were used for evaluating the accuracy of the classifier [2].

The classifier module used in the tool is based on Naïve Bayes model. It involves a simplifying conditional independence assumption. This assumption does not affect the accuracy in text classification by much but makes really fast classification algorithms applicable for the problem. Rennie et al discuss the performance of Naïve Bayes on text classification tasks in their 2003 paper [2] [5]. The code of the analysis tool is available at [17]. Training involved preprocessing data and applying negation handling before counting the words. Since we were using Bernoulli Naive Bayes, each word is counted only once per document. On a laptop running an Intel Core 2 Duo processor at 2.1 GHz, training took around 1 minute 30 seconds and used about 700 megabytes of memory. The memory usage stems largely from bigrams and trigrams prior to feature selection [2].

The outcome was that the tool obtained an overall classification accuracy of 88.80% on the test set of 25000 movie reviews [2].

THE TEXT USED FOR ANALYSIS AND THE GOAL OF ANALYSIS

The goal of analysis was to determine the correlation between text sentiments and engagement levels. The hypothesis was as follows.

- The texts are actual posts which have the goal to help people with mental disorders. The community involved are people who among other tags follow the tag “mental health” and “love”. All texts are produced by one author, Tomislav Krištof who has the access to posts data and statistics [7].
- The goal was to prove the hypothesis which states that the text of positive sentiment should be more engaging than the texts of negative sentiment to people who seek education, encouragement, guidance and help regarding mental health in a therapeutic sense.
- The engagement is measured by number of recommendation each post gets. To equalize the recommendation measurement, the read-recommendation ratio was used. The read-recommendation ratio is a quotient between number of recommendation and reads.

$$Read - recommendation\, ratio = \frac{Number\ of\ recommendations}{Number\ of\ reads} \qquad (1)$$

- Higher ratio means more engaging the post.
- Posts include 24 texts about mental health and love and are aimed to help people in need. Total number of words is 15199, with the average of 633 words. The median is 431 words.
- Total number of unique views is 1674, and total number of reads is 1157. The total number of reads is considered as a sample size, (*n* = 1157). The read is determined by the platform and is calculated by *unknown* variables (not publicly available). The workaround to find out is to determine the way Medium.com knows when the text has been scrolled to the end. Using some JavaScript, it is possible to capture an event that fires when text is scrolled to the end [18] [19]. However, in good faith it is probable that since the Medium.com has the algorithm that calculates the time necessary to read the post, it also uses that variable in addition to scroll to distinguish between read posts and fast scrolled posts.

SENTIMENT ANALYSIS

THE CORRELATION BETWEEN THE SENTIMENT AND READ – RECOMMENDATION RATIO

To determine the correlation between the sentiment of posts, each post was analyzed for sentiment via mentioned tool [1]. Each post was given sentiment status as shown in the following example (9 rows are shown of 24):

Table 1: Quantification of the sentiment

1	2	3	4
Sentiment (P/N)	Sentiment (100/-100)	Probability of (2) (0-100)	Pondered Sentiment
P	100	100,00	10,00
N	-100	100,00	-10,00
N	-100	73,76	-7,38
P	100	100,00	10,00
P	100	100,00	10,00
N	-100	81,10	-8,11
P	100	100,00	10,00
P	100	100,00	10,00
P	100	99,37	9,94

First the positive sentiment is given the value of 100 and the negative sentiment the value of -100. Than the probability was taken as a ponder (probability of the sentiment determination). And as a result, the Pondered sentiment is determined with the formula

$$Pondered\ sentiment = \frac{Sentiment \cdot Probability}{1000} \qquad (2)$$

After quantifying the sentiment, every post was given a Read-recommendation ratio (column 6 in Table 2.).

Table 2: The whole sample in numbers

1	2	3	4	5	6	7
Post no.	Views	Reads	Read ratio	Recom-mends	Rec.-Read ratio · 100	Pondered sentiment
18	51	31	60,78	10	32,26	10,00
22	32	27	84,38	11	40,74	-10,00
6	181	130	71,82	18	13,85	-7,38
17	16	11	68,75	3	27,27	10,00
23	11	10	90,91	5	50,00	10,00
13	149	79	53,02	23	29,11	-8,11
19	10	4	40,00	2	50,00	10,00
7	49	27	55,10	5	18,52	10,00
24	21	21	100,00	11	52,38	9,94
21	16	16	100,00	5	31,25	-10,00
15	50	49	98,00	9	18,37	9,60
8	19	17	89,47	2	11,76	-10,00

20	70	59	84,29	15	25,42	-9,99
16	38	32	84,21	7	21,88	10,00
14	77	64	83,12	12	18,75	9,90
4	356	293	82,30	33	11,26	-10,00
3	11	9	81,82	1	11,11	-9,90
2	121	89	73,55	7	7,87	-10,00
10	75	53	70,67	10	18,87	9,69
5	131	63	48,09	13	20,63	9,99
12	26	12	46,15	4	33,33	10,00
11	20	9	45,00	3	33,33	-10,00
9	58	22	37,93	7	31,82	-8,20
1	86	30	34,88	4	13,33	-10,00

Using the Pearson's correlation coefficient, it was easy to calculate the correlation between column 6. and column 7.

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}} \qquad (3)$$

where:

- n, x_i, y_i are defined as above
- $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ (the sample mean); and analogously $\overline{y}$

In the above (2) fomula [21] we entered values (column 6., as x, and column 7., as y) and calculated the Pearson's correlation coefficient r.

The result:

$$r = 0{,}340038418$$

Under classification according to Cohen, J. [20] we get the following determination:

Table 3. Classification of Pearson's correlation coefficient r

Coefficient Value	Strength of Association
0.1 < | r | < .3	small correlation
0.3 < | r | < .5	medium/moderate correlation
| r | > .5	large/strong correlation

So, we can conclude that there is moderate positive correlation between the Post sentiment and Read – recommendation ratio based on the sample size of *n* = 1157., where n is number of unique reads.

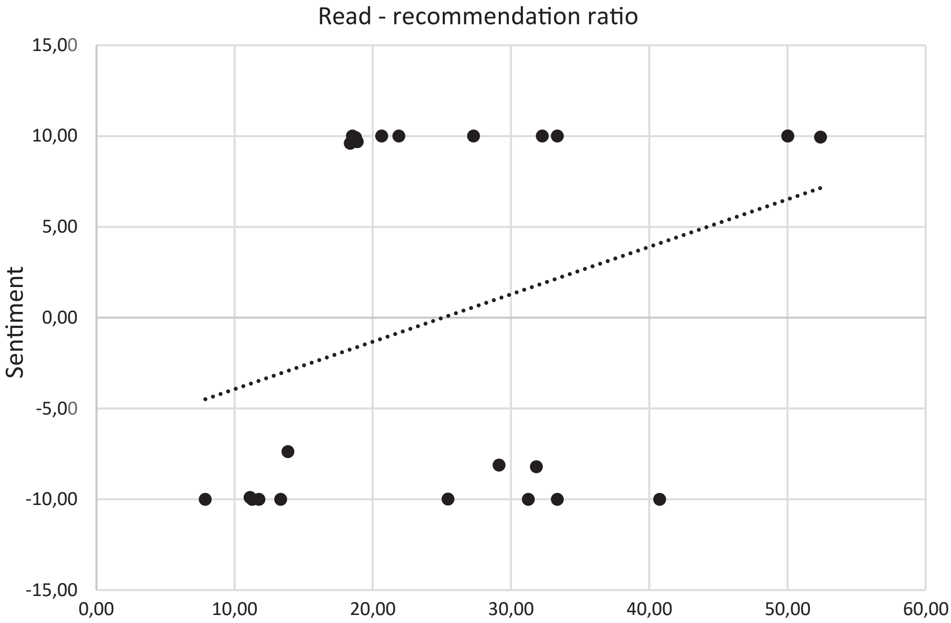

Figure 1: The correlation between Sentiment and Read – recommendation ratio

On the Figure 1. we can see the apparent skewness that nicely depicts the correlation between the sentiment and read – recommendation ratio. The “trend line” was added for better depiction.

THE CORRELATION BETWEEN USING THE WORD “LOVE” IN THE POSTS AND READ RATIO

Furthermore, we made another hypothesis which states, there can be a correlation between using the word “love” and read ratio.

Table 4. Word density “love” and read ratio

1	2	3	4	5	6
Word density „love“	**Views**	**Reads**	**Read ratio**	**No. of word „love“**	**Post lenght (no. words)**
2,85326087	51	31	60,78	21	736
1,639344262	32	27	84,38	6	366
1,408450704	181	130	71,82	5	355
1,339285714	16	11	68,75	6	448
0,892857143	11	10	90,91	2	224
0,71942446	149	79	53,02	5	695
0,260416667	10	4	40,00	3	1152
0,156128025	49	27	55,10	2	1281

Word density “love” is calculated by the following formula:

$$Word\ density\text{"love"} = \frac{No.of\ word\text{"}love\text{"}}{Post\ lenght\ in\ words}$$

Read ratio is calculated by the following formula:

$$Read\ ratio = \frac{Reads}{Views}$$

By using the Pearson’s correlation coefficient, it was convenient to calculate the correlation. The result was: r = 0,3447607. To calculate it we use the columns 1. and 4. of the Table 4. Note 1: the calculation is based on 319 reads of unique readers.

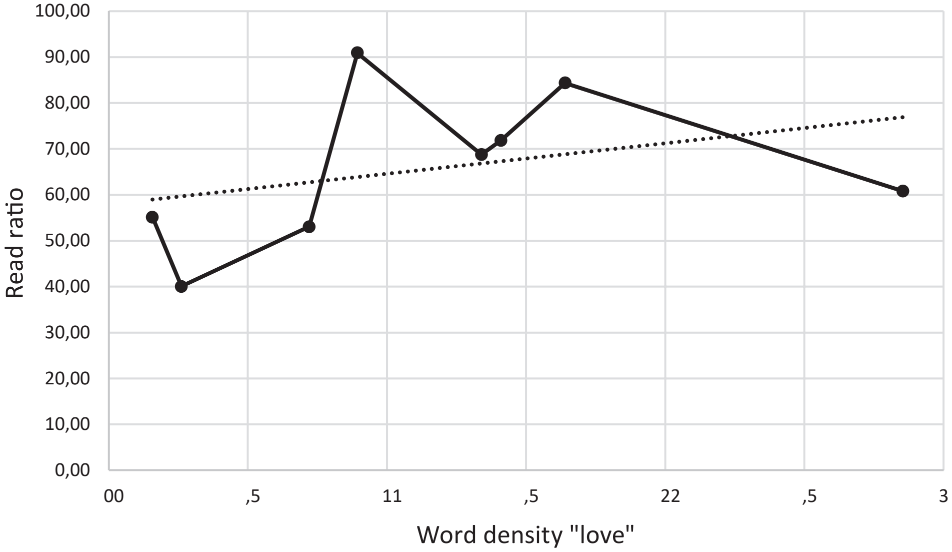

Figure 2: The correlation between word density “love” and read ratio

SENTIMENT TITLE VS. READ RATIO

Based on general opinion that negative or slightly negative titles of articles provoke more clicks and interaction we decided to test that hypothesis. We chose 11 titles and 632 reads (n=632) to prove the hypothesis: Negative titles provoke engagement. We have calculated the sentiment of the content titles and made a correlation analysis between the content title sentiment and read ratio.

$$Read\ ratio = \frac{Reads}{Views}$$

We discovered that the r=-0,241740106, which determines that there is a small negative correlation between the title sentiment and read ratio. It can be described as follows. The visitors of the content are more likely to read posts which bear titles of negative sentiment. The data is presented in Table 5.

Table 5. Title sentiment and read ratio

1	2	3	4
Views	**Reads**	**Read ratio**	**Title sentiment**
51	31	60,78	-100
32	27	84,38	-100
11	9	81,82	-100
75	53	70,67	-100
131	63	48,09	-100
10	4	40,00	100
38	32	84,21	100
356	293	82,30	100
121	89	73,55	100
20	9	45,00	100
58	22	37,93	100
903	**632**		

CONTENT LENGTH VS READ RATIO

It is no coincidence that the viewers are more likely to read shorter content. On the sample of 1674 views and 1157 reads (*n* = 1157), we were able to prove that there is a strong correlation between post size in words and read ratio. The correlation coefficient is *r* = –0,85111966. This proves strong negative correlation between posts length and read ratio.

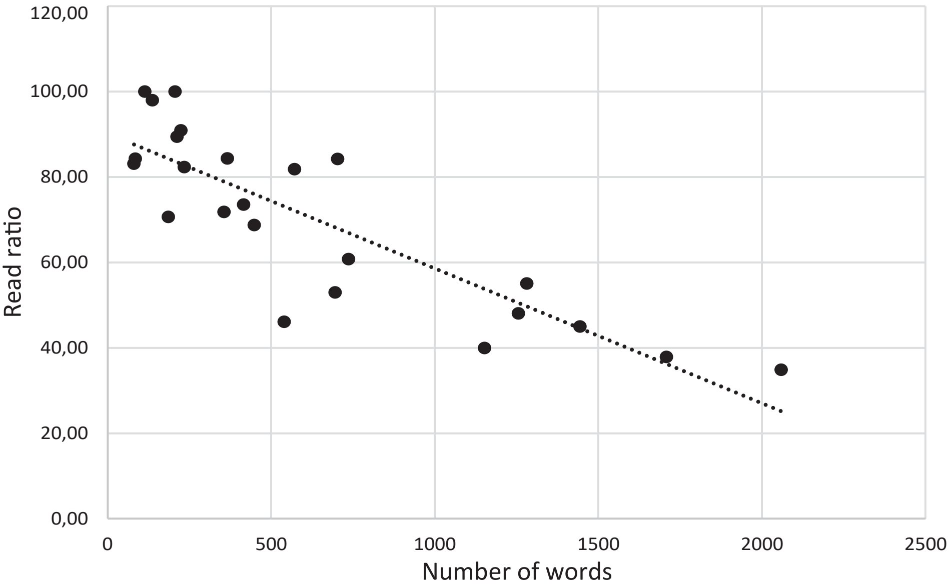

Figure 3: The correlation between read ratio and number of words in posts

POSITIVE TREND OF RECOMMEND READ RATIO

We have discovered also that there is a moderate correlation between post history and recommend ratio. The newer the post is, the more likely is that it will be recommended. This can be interpreted in following frame. As the time progresses the author of the posts learns what the readers like, and writes appropriate content. Also as time progresses there are more followers of the author, so the social proof element also influences the decision to recommend. Also, the number of returning followers who like the author increments with time, and that number of returning followers can only be higher, never lower comparing to the beginning of posting.

Sample consists of 1674 views and 1157 reads, (*n* = 1157). The calculated *r* = 0,381809719 proves moderate correlation.

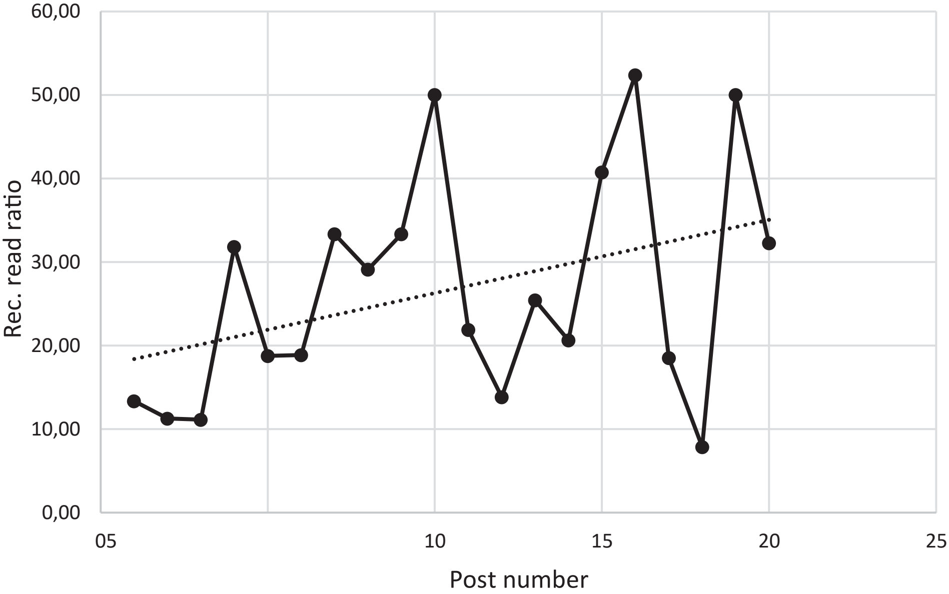

Figure 4: The correlation between rec. read ratio and post number

CONCLUSION

We can conclude that there are many correlations to be found regarding sentiment and format of the content published on social network Medium.com and that from these correlations we can learn and predict the engagement for each post. Our predictions are based on correlation analysis with different results, but within all results there are at least minimum correlation to each model. Therefore, we are able to conclude as follows:

1. There is a correlation between the sentiment and read – recommendation ratio. The correlation is moderate, *r* = 0,340038418. The conclusion is: If we wrote a post with positive sentiment we are more likely to receive recommendations.
2. There is a correlation between using the word "love" in the posts and read ratio. The correlation is moderate, *r* = 0,3447607. The conclusion is: If we use the word love within appropriate density, we can expect the readers to read the content through. Variation of word density ("love"), is (0,15 – 0,26).
3. There is a small negative correlation between the title sentiment and read ratio. Negative sentiment in the title provokes more readings. The r=-0,241740106 proves small but present correlation between those two variables. The conclusion is: If we make the title negative, there is more chance of reading the post through
4. Thereisastrongnegativecorrelationbetweenthecontentlengthandreadratio. The *r* = –0,85111966 proves strong negative correlation. The conclusion is: The readers are more likely to read shorter posts through.
5. There is a moderate correlation between the post number and read ratio. The *r* = 0,381809719 proves moderate correlation. The conclusion is: As the time progresses posts get more and more recommendations. This can be explained by the learning curve of the author to produce more appropriate content for the readers.

REFERENCES

[1] (2017-04-20) http://sentiment.vivekn.com/

[2] Narayanan V., Arora I., Bhatia A. (2013). Fast and accurate sentiment classification using an enhanced Naive Bayes model. Intelligent Data Engineering and Automated Learning IDEAL 2013 Lecture Notes in Computer Science Volume 8206, 2013, pp 194-201

[3] (2017-05-20) http://text-processing.com/demo/sentiment/

[4] (2017-05-20) http://ai.stanford.edu/~amaas/data/sentiment/

[5] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. (2011). Learning Word Vectors for Sentiment Analysis. The 49th Annual Meeting of the Association for Computational Linguistics (ACL 2011).

[6] (2017-05-20) http://www.imdb.com/

[7] (2017-05-20) https://medium.com/@tomo.kristof

[8] Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. "Thumbs up?: sentiment classification using machine learning techniques." Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10. Association for Computational Linguistics, 2002.

[9] (2017-05-20) https://medium.com/tag/mental-health

[10] (2017-05-20) https://medium.com/tag/love

[11] Chowdhury, G. (2003) Natural language processing. Annual Review of Information Science and Technology, 37. pp. 51-89. ISSN 0066-4200,
(2017-05-20) http://dx.doi.org/10.1002/aris.1440370103

[12] (Bangalore & Joshi, 1999; Barker & Cornacchia,2000; Chen & Chang, 1998; Dogru & Slagle, 1999; Kam-Fai et al.. 1998; Kazakov et al.. , 1999; Lovis et al.. 1998; Tolle & Chen, 2000; Zweigenbaum & Grabar, 1999)

[13] (Kehler, 1997; Mihalcea & Moldovan,1999; Meyer & Dale, 1999; Pedersen & Bruce, 1998; Poesio & Vieira,1998; Tsuda & Nakamura, 1999)

[14] (Argamon et al.., 1998; Fernandez & Garcia-Serrano, 2000; Martinez et al.., 2000, 1998).

[15] Cambria, E; Schuller, B; Xia, Y; Havasi, C (2013). “New avenues in opinion mining and sentiment analysis”. IEEE Intelligent Systems. 28 (2): 15–21. doi:10.1109/MIS.2013.30.

[16] Stevenson, Ryan; Mikels, Joseph; James, Thomas (2007). “Characterization of the Affective Norms for English Words by Discrete Emotional Categories” (PDF). Behavior Research Methods. 39 (4): 1020–1024. doi:10.3758/bf03192999. PMID 18183921.

[17] (2017-05-20) https://github.com/vivekn/sentiment

[18] (2017-05-20) https://medium.com/p/74b9f41509b/state/read

[19] (2017-05-20) https://www.quora.com/How-does-Medium-determine-whether-an-article-has-been-read

[20] Cohen, J. (1988) Statistical Power Analysis for the Behavioral Sciences, 2nd ed. Hillsdale, NJ: Erlbaum.

[21] (2017-05-20) https://en.wikipedia.org/wiki/Pearson_correlation_coefficient

[22] (2017-05-20) http://www.alexa.com/siteinfo/medium.com