International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 2, Issue 5, pp 28-32, July-2018

Statistic-Based Sentiment Analysis of Social Media Data

[1]Aman Karn, [2]Anish Shrestha, [3]Anil Pudasaini, [4]Binay Mahara, [5]Anku Jaiswal

[1,2,3,4]Student, Advanced College of Engineering and Management, Kupondole, Kathmandu, Nepal / Tribhuvan University, Kritipur, Nepal

[5]Lecturer, Computer and Electronics Engineering Department, Advanced College of Engineering and Management, Kupondole, Kathmandu, Nepal

***Abstract* - This paper presents one of the practices of text/opinion mining of web data which can help to provide assistance to prepare reports for Customer Relationship Management (CRM). For convenience we use Twitter tweets data for sentiment analysis. These sentiment values are further treated with statistical evaluations like z-tests and chi-squared test. As social media data are considered as normally distributed, the test like ANOVA test, t-test (for small sample-size) and hypotheses test can be used. Python programming language is used for the task as it has several libraries and packages for Natural Language Processing, statistics, data visualization while supporting the features of general-purpose programming language. This paper also dictates the process of fetching, storing, cleaning, language - translating, sentiment & statistical evaluating, simulating & building theoretical model for hypothesis testing of the data.**

Keywords: Sentiment Analysis, Natural Language Processing (NLP), Twitter Sentiment, Opinion Mining, Tweets analysis, Chi-squared test, Z-test, Social media data visualization.

I. INTRODUCTION

Sentiment analysis is the form of text/opinion mining using natural language processing, computational linguistics and text evaluation to extract quantifiable information from piece of text corpus to deduce writer's intention towards the adjective used in that corpus and categories positive, negative and other levels. This sentiment analysis report is further used in Hypothesis Testing for the independence test and goodness-of-fit test.

Sentiment values are extracted based upon the phrases of adjectives and adverbs. Adjectives are considered to be good indicators of sentiment in the sentence. So acquiring the knowledge of semantic orientation [1] of sentence can be beneficial for polarity determination of sentence.

II. RELATED WORKS

Peter D. Turney, (2002) [2], presented an unsupervised learning algorithm which is used to classify reviews as recommended (thumbs up) or not recommended (thumbs down). The algorithm proceeds by predicting the semantic orientation of phrase in given review that contain adjectives or adverbs or their combinations. The model was formulated on reviews of automobiles, banks, movies and travel destination.

According to their method, primary task is to extract phrase containing adjectives or adverbs. Former works in Sematic Orientation (Hatzivassiloglou&Wiebe et al., 2001) [1], indicate that adjectives can act as good indicators of subjectivity in evaluation of sentences. From his report, an isolated adjective may indicate subjectivity, and there may be insufficient context to determine semantic orientation. For example, the adjective "unpredictable" may have a negative orientation in an automotive review, in a phrase such as "unpredictable steering", but it could have a positive orientation in a movie review, in a phrase such as "unpredictable plot". So, two consecutive words are extracted where the one member is an adjective or adverb and second provides context. Word pairs are extracted form review only when any of patterns matched form table below:

First Word	Second Word	Third Word (Not Extracted)
JJ	NN OR NNS	Anything
RB, RBR or RBS	JJ	Not NN nor NNS
JJ	JJ	Not NN nor NNS
NN OR NNS	JJ	Not NN nor NNS
RB, RBR, or RBS	VB, VBD, VBN or VBG	Anything

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 2, Issue 5, pp 28-32, July-2018

The Part-of-Speech (POS) tag list:

JJ	adjective	e.g. big
NN	noun, singular	e.g. desk
NNS	noun plural	e.g. desks
RB	adverb	e.g. very, silently
RBR	adverb, comparative	e.g. better
RBS	adverb, superlative	e.g. best
VB	verb, base form	e.g. take
VBD	verb, past tense	e.g. took
VBG	verb, gerund/present Participle	e.g. taking
VBN	verb, past participle	e.g. taken

The second step involves calculating of Semantic Orientation using algorithm, called PMI-IR [2], uses Point wise Mutual Information (PMI) and Information Retrieval (IR) to measure the similarity of pairs of words (Church & Hanks, 1989). PMI-IR is empirically evaluated using 80 synonym test questions from the Test of English as a Foreign Language (TOEFL) and 50 synonym test questions from a collection of tests for students of English as a Second Language (ESL). On both tests, the algorithm obtains a score of 74%. PMI-IR is contrasted with Latent Semantic Analysis (LSA), which achieves a score of 64% on the same 80 TOEFL questions.

PMI between two words, word1 and word2 is calculated as:

$$\mathrm{PMI}(\mathrm{word}_1, \mathrm{word}_2) = \log_2\left(\frac{\mathrm{p}(\mathrm{word}_1 \text{ and } \mathrm{word}_2)}{\mathrm{p}(\mathrm{word}_1)\mathrm{p}(\mathrm{word}_2)}\right)$$

Here, p(word1 & word2) is the probability that word1 and word2 co-occur. If the words are statistically independent, then the probability that they co-occur is given by the product p(word1) p(word2). The ratio between p(word1 & word2) and p(word1) p(word2) is thus a measure of the degree of statistical dependence between the words. The log of this ratio is the amount of information that we acquire about the presence of one of the words when we observe the other. The Semantic Orientation (SO) of a phrase, is calculated here as follows:

SO(phrase) = PMI(phrase, "excellent") – PMI(phrase, "poor")

The reference words "excellent" and "poor" were chosen because, in the five-star review rating system, it is common to define one star as "poor" and five stars as "excellent". SO is positive when phrase is more strongly associated with "excellent" and negative when phrase is more strongly associated with "poor".

III. DATA COLLECTION AND SENTIMENT ANALYSIS

Python programming language is used for the task because of its simplistic syntax and open libraries and large variety of forum and community for help. In addition, python is quite powerful with the stuff of numeric computation, natural language processing, artificial-intelligence, etc. The package tweepy-api helps to stream data (tweets) from Twitter with the help of proper keys and tokens. For data cleaning, Regex (Regular Expression package) is used to clean tweets from HTML tags and codes. For further multiple analysis, storing the tweets data seems appropriate. JSON serialization was found to be efficient at handling the task. The JSON format of storing the data is similar to dictionary-datatype of python. So python can handle and operate on the tweets as working with its own data structure. Also the JSON file format is compatible with many other programs. For the purpose of Sentiment Analysis, Textblob package is used. Textblob is NLP package build on top of NLTK package. Textblob can handle all the tasks of NLTK like word/sentence tokenization, part-of-speech tagging, stemming, noun-phrase extraction, parsing with addition features like sentiment analysis, classification, language detection & translation powered-by Google translate, WordNet integration, etc. The tweets data is fed in Naïve-Bayes Classifier which has been trained on the Sentiment Polarity Lexicon from WordNet with preprocessing for classifier training.

Each word in lexicon has score for:

Polarity: Negative ‣ Positive	(-1 ‣ +1)
Subjectivity: Objective ‣ Subjective	(+0 ‣ +1)
Intensity: modify next word	(x0.5 ‣ x2.0)

First the language of tweet is detected. If it confirms to be in English language, sentiment extraction task is carried out. If the language is other than English, at first it translated to English and then sentiment is extracted. Since extensive amount of work has to been done in building WordNet in English, the sentiment extraction is only effectively possible for English tweets. If sentiment is extracted for foreign language tweets, classifier doesn't recognize it and classifies them as neutral value which can considered to be flaw of Naïve-Bayes classifier.

The nature of contemporary tweets (or as general in, social media opinions) are quite diverse due to inclusion of foreign-language, emoticons, slang, sarcasm, and mixed languages. Tasks like sentiment analysis, gist understand, etc. are not an easy task. So the insights of the Statistics and Probability can shed new light to the observation.

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 2, Issue 5, pp 28-32, July-2018

IV. STATISTICAL ANALYSIS AND DATA VISUALIZATION

According to Central Limit Theorem, with larger number sample size (n>50), any distribution with finite variance have sample mean of distribution approximately normal. And our analysis and calculation is based upon the fact that the samples are normally distributed.

The following is the statistic measure of a day tweet fetched for the word "Google" on 2018/06/02:

TABLE I
Statistic measure of "Google" on 2018/06/02

Statistics	All data	Not Neutral data
No. of Observation	10010	5366
Sentiment range	[-1, 1]	[-1,1]
Mean	0.070245	0.131038
Variance	0.078869	0.139172
Skewness	0.010119	-0.479421

Observing the opinion (positive, negative & neutral) distribution while comparing with distribution binary value (only positive & negative).

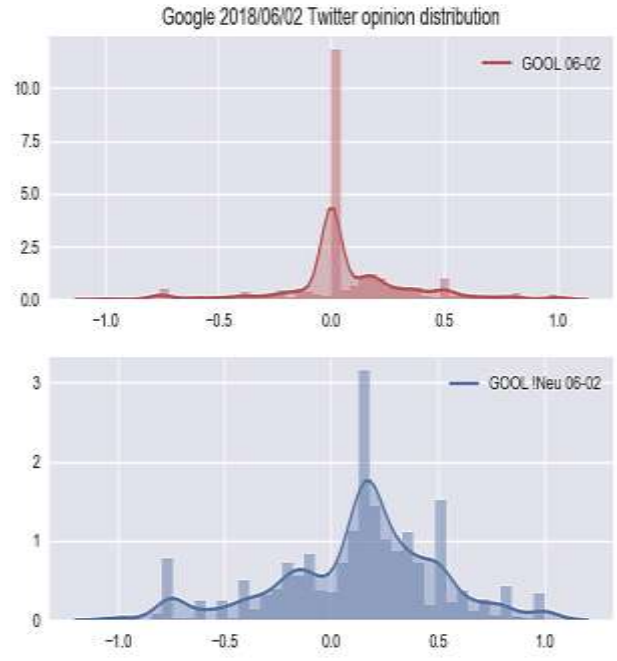

Figure-1: All opinion distribution (top) and Excluded neutral distribution (bottom)

Again comparing these two data in pie chart:

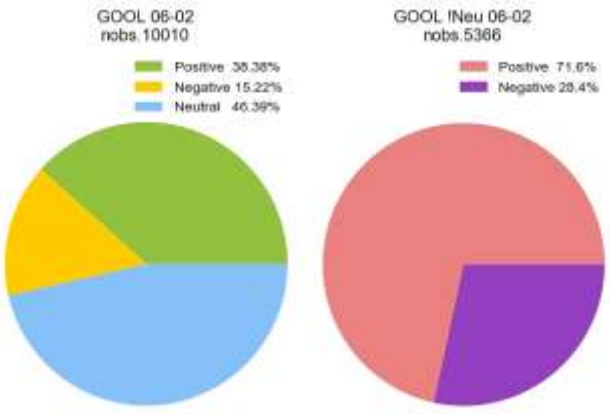

Figure-2: All opinion pie chart (left) and Excluded neutral distribution (right)

Conducting the Z test (proportion) for the binary value (only positive & negative) data for observation from 2018 June month with hypothesis as formulated:

Null hypothesis H0: Positive and Negative are equal in numbers. P = 0.5i.e. Value = 1 or 'PASS'

Alternative hypothesis H1: Positive and Negative are not equal in numbers. P ≠ 0.5 i.e. value = 0 or 'FAIL'

The results are as follows:

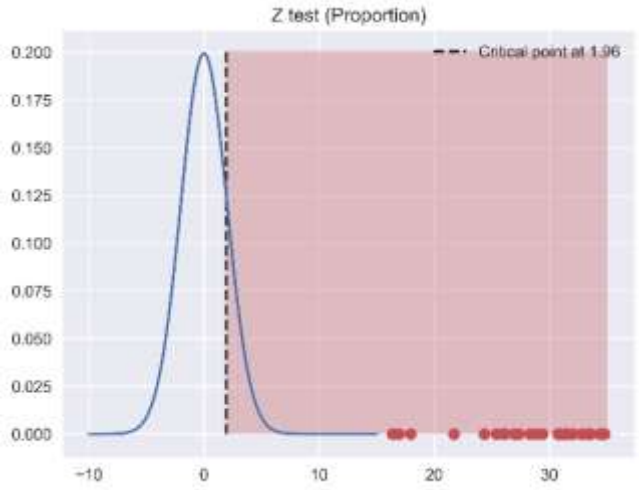

Figure-3: Z-test (proportion) for positive and negative opinions

The critical point for 5% significance level for Two-tailed test = 1.96. The red-dots are the computed Z scores (which are greater than 1.96) lies in rejected region. So, the null hypothesis is rejected and alternative hypothesis is accepted. It can be concluded that the numbers of positive and negative opinion are not equal.

Observing the means of sentiment scores with their confidence interval we found following results:

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 2, Issue 5, pp 28-32, July-2018

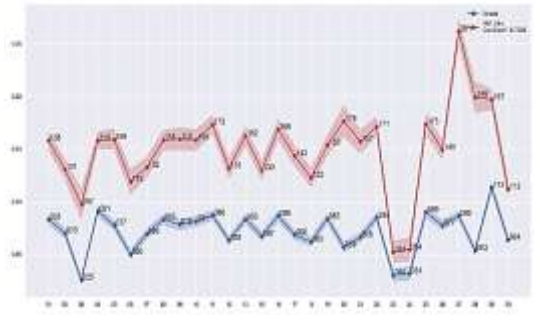

Figure-4: Excluded neutral means values (Red-top) and all opinion (Blue-bottom)

The solid line denotes the mean line and the corresponding faded shade shows confidence (95%) region The Pearson Correlation Coefficient between above result found to be 0.7235. So the nature of test result shown by excluded neutral data is similar to nature of test result of whole data.

Separating the positive, negative and neutral sentiment score and observing for all 2018 June samples:

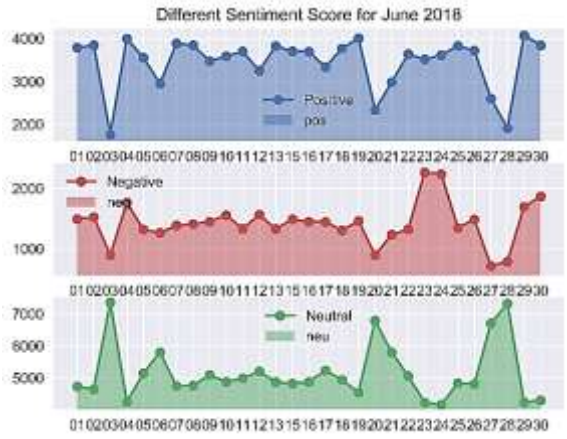

Figure-5: Line graph (Top-Blue) Positive, (Middle-Red) Negative and (Bottom-Green) Neutral sentiment

Also in heatmap:

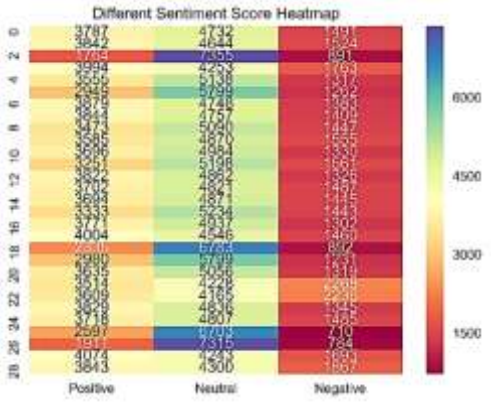

Figure-6: Heatmap (Right) Positive, (Middle) Neutral and (Bottom-Green) Negative with sentiment score

For further calculation, we tried to simulate the data using its sample size, mean, and standard deviation.

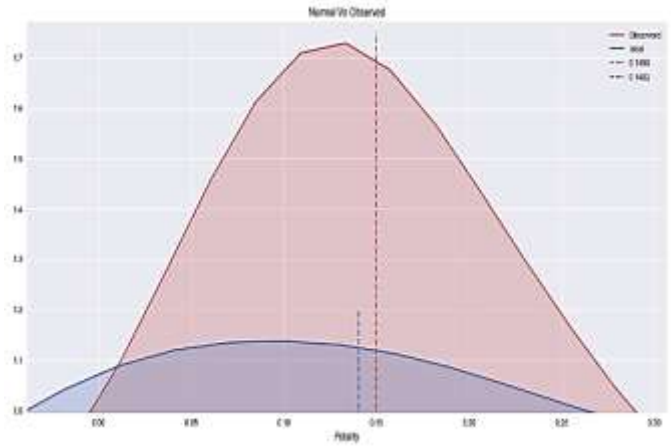

Figure 7: Actual (Red) VS simulated (Blue) distribution of samples

For example: actual distribution mean = 0.1499

Simulated distribution mean = 0.1402

For consistent and uniform calculation, we kept the random seed value same for all other observations. Following figures shows the actual and simulated distribution for few of observations:

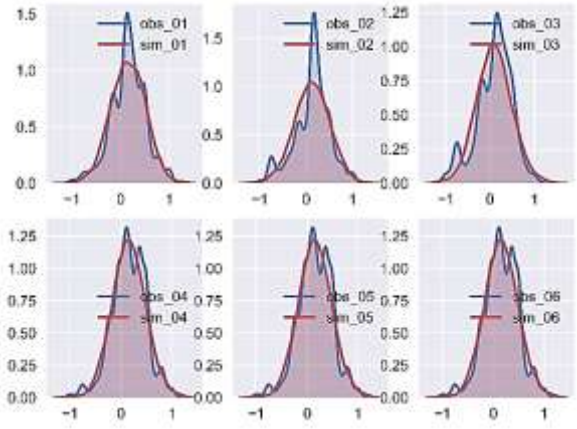

Figure-8: Actual (Blue) and Simulated (Red) distribution for first six samples

Similarly, simulation for rest of the data is carried out. These simulated results are used in Chi-Squared test. The hypothesis is:

Null hypothesis H0: Opinions are fare.

Observed and Expected are same.

Alternative hypothesis H1: Opinions are not-fare (biased).

Observed and Expected are different.

Observed opinion are the actual data and the Expected values are the randomly generated simulated data.

ISSN (online): 2581-3048
Volume 2, Issue 5, pp 28-32, July-2018

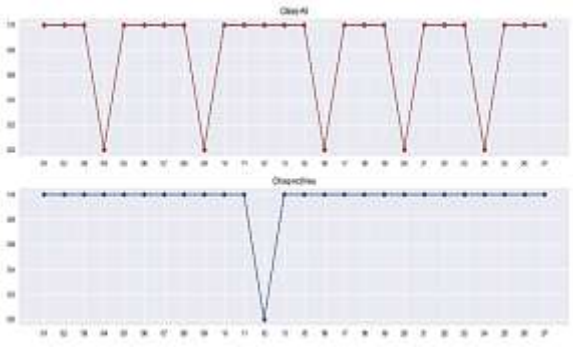

Figure-9: Chi-Squared test results. Pass (1) and Fail (0). (Top) whole data test. (Bottom) Neutral Excluded data

The (Red-top) graph shows the Chi-Squared test results into level, 1-pass and 0-fail. In Whole data, 5 samples out of 29 fail the test. In (Blue-bottom), for Neutral excluded, 1 out of 29 failed the test.

So we can conclude that majority of sample passed the Chi-Squared test. Therefore, we accept the null hypothesis as Opinion/Sentiment for term "Google" for month 2018 June is fare.

Later Analysis of Variance (ANOVA) test was carried out to check the difference on data.

The hypothesis is:

Null hypothesis H0: Opinions are not significantly different.

Alternative hypothesis H1: Opinions are significantly different.

The test was carried out for all samples. The result:

Since

$F_{cal} = 54.752 > F_{tab}(0.05)(28,141187) = 1.4764$

Therefore, H0 is rejected and H1 is accepted.

V. CONCLUSION

Thus this statistical-based sentiment analysis helps to draw out the quantifiable results from mundane opinion expressed in social media. This utilize of statistical hypothesis testing method to reach a conclusion about the population distribution towards a particular topic/subject of interest based upon the samples gathered from social media/web provides vital information for further processing, model building and evaluation as empirical evidence.

REFERENCES

[1] Hatzivassiloglou, V. and McKeown, K. (1997). Predicting the semantic orientation of adjectives. Proceedings of the 35th annual meeting on Association for Computational Linguistics -. [online] Available at: *https://dl.acm.org/citation.cfm?id=979640. DOI: 10.3115/976909.979640.*

[2] Turney, P. (2001). Thumbs up or thumbs down? Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02. [online] Available at: *https://dl.acm.org/citation.cfm?id=1073153. DOI: 10.3115/1073083.1073153.*

[3] Pagolu, V., Reddy, K., Panda, G. and Majhi, B. (2016). Sentiment analysis of Twitter data for predicting stock market movements. 2016 International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES). [online] Available at: *http://ieeexplore.ieee.org/document/7955659/?reload=true.DOI: 0.1109/SCOPES.2016.7955659.*

[4] Fang, X. and Zhan, J. (2015). Sentiment analysis using product review data. Journal of Big Data, [online] 2(1). Available at: *https://link.springer.com/article/10.1186/s40537-015-0015-2. DOI 10.1186/s40537-015-0015-2.*

[5] Church, K. and Hanks, P. (1989). Word association norms, mutual information, and lexicography. Proceedings of the 27th annual meeting on Association for Computational Linguistics-. [online] Available at: *https://www.researchgate.net/publication/2477223_Word_Association_Norms_Mutual_Information_and_Lexicography.DOI: 10.3115/981623.981633.*

[6] Bird, S., Klein, E. and Loper, E. (2011). Natural language processing with Python. Beijing [etc.]: O'Reilly. *ISBN:0596516495 9780596516499.*

Citation of this article:

Aman Karn, Anish Shrestha, Anil Pudasaini, Binay Mahara, Anku Jaiswal, "Statistic-Based Sentiment Analysis of Social Media Data", *International Research Journal of Innovations in Engineering and Technology (IRJIET)*, Volume 2, Issue 5, pp 28-32, July 2018.

****

