iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 8 **Issue:** V **Month of publication:** May 2020

DOI: http://doi.org/10.22214/ijraset.2020.5119

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue V May 2020- Available at www.ijraset.com

Sentiment Analysis on Youtube & Twitter Data using Machine Learning

Prof Bhajibhakare M. M.[1], Ankita Borkar[2], Simran Naik[3], Suvarna Solase[4], Padmasen Kunjir[5]
[1, 2, 3, 4, 5]*Department of Computer Engineering, JSCOE, Pune, SPPU*

Abstract: Sentiment analysis is very recent trend in Web intelligence and information extraction. Now a days analysis of text is depend upon dictionary of emotions also depends upon creation of emotion dictionary, artificial design and mining from that dictionary. In this paper for classification and analysis we using Naïve Bayes and Support Vector Machine (SVM) techniques are used.
Keywords: Web intelligence; Extraction; Deep Learning; SVM; Neural Network

I. INTRODUCTION

A. The Motive

We have decided to work with social media (Twitter) because we feel that it is a better reflection of social sensitivity in comparison to standard online articles and various web blogs. The magnitude of relevant information is greater on twitter, as compared to traditional blogging sites. In addition the response is also downloadable and is also general. Recent conceptual analysis in modern life as analyzing the stock market data of a particular organization. This can be done by examining the general public opinion of the organization by looking at time and applying economic strategies to find the link between the public sentiment and the stock market value of the organization. An organization can also determine how well their product responds to today's market, which market segments have a positive reaction and a negative response. If an organization has access to this information then it can analyze the causes behind the different reactions, so it can begin to present its product in a revised way to get better results such as developing good market places.

II. WORK DONE

The word bag model is an important model for text segmentation because it is easier to manage and work better. It represents a text to be categorized as a fund or a set of words without a link / trust of one word, e.g. It does not focus on the language and word order within the text. This model was well received by investigators. We can accept this model in our classifier using unigrams. Usually speaking n-grs is a sequence of words "n" of text we have. Unigrams are thus used to group each word into a text to be distinguished, and we can predict the probability of a one-word love event with one another in the presence or absence of a text. This is a very simple prediction but has been shown to work better. The easiest way to use unigrams is to assign the original unity to it, and to consider the ratio of the whole text, and it is done by adding all the paths of all unigrams. Before the pronunciation of a word can be good if the word is subjective, think of the word “sweet”; it can be a bad thing if the word usually comes with bad tones, Think of the word "scary" it can have a strong following and the word "respect" has a good past tense but with a weak submission in certain cases.

We can mention three instances of using pre-color as symbols. An easy-to-control condition for using public dictionaries / dictionaries online that displays the name on its previous map. The Multi-Perspective-Question-Answering (MPQA) is an online resource with low rankings that refers to "good" or "negative" and "strong" or "weak". SentiWordNet is one that offers the possibility of all the names of good, bad, and neutral partners [15]. The second condition is to build a previous dictionary to distinguish the unity from our trained data in terms of the presence of words across distances. This situation has been shown to provide high performance, since previous word coherence is suitable for a particular text type. However, the last one is a cautious approach because the training information should be labeled in the appropriate sections before calculating the word availability for each category. Kouloumpis et al. note the decrease in performance using the lexicon names and n-gram custom namespaces generated from the training data, separately when using n-gr alone [7].

The third scenario is the world between the two scenarios above. In this case we are creating our own polarity lecyon but not with our training data, so we don't need to have label inserts.

Most investigators in this field have used the existing old dictionary dictionary (e.g., [7], [12] and [16]) while many others have also examined building their own polarity dictionaries (e.g., [3]], [10] and [11]).A basic problem about the former method of maturation

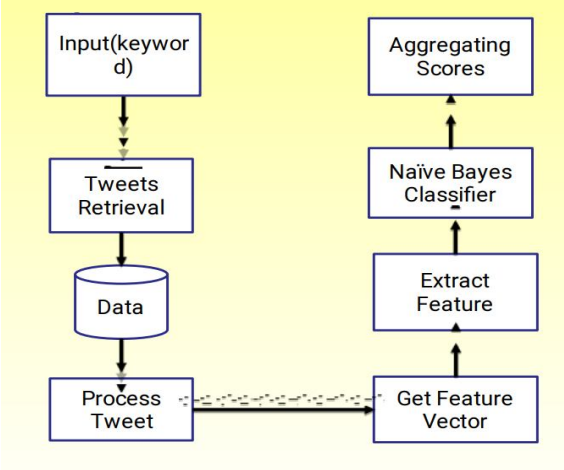
has been identified by Wilson et al. they distinguish between the previous polarity and the position polarity [16]. They say that the original word cohesion may actually be different than the way the word is used in a particular context. The paper introduced the following phrase as an example:

Philip Clapp, president of the National Environment Trust, summarizes the whole point of the environmental response: "There is no reason to believe that these polluters will later make sense."

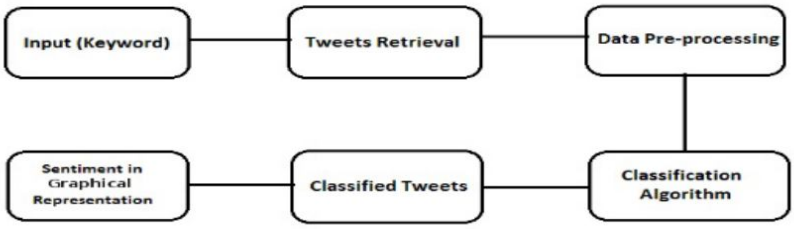
In this example all four terms listed as “trust”, “well”, “reasoned” and “reasonable” have positive aspects when viewed outside the context of a sentence, but here they are not used to express the concept. This concludes that while saying the word "trust" may be used in positive sentences, but this does not allow for the possibility of non-constructive sentences.

III.METHODOLOGY

A. *Design Diagram*



B. *Proposed-System*



C. *Feature-Extraction*

Now that we have arrived at our training set we need to extract some useful features in it that can be used in classification. But first we will discuss some text formatting techniques that will help us to extract the feature:

1) *Marking:* It is a way of dividing a line of text into words, symbols and other logical things called "tokens". The tokens can be separated by whites and / or character characters. With the help of that, it can go through tokens just like every other component that creates a tweet.
2) Url and user references (denoted by "http" and "@" tokens) are removed if we are interested in analyzing only the tweet text.
3) The punctuation and numbers / numbers can be removed if for example we wish to compare t list with English word list.
4) *Low Conversion:* Tweet can be customized by converting it into a dashboard that compares to an English dictionary.
5) *Elongation:* It is an effective process of reducing the word from its roots or stem [28]. For example, the stem stem has reduced the terms “stemmer”, “stemmed”, “stemrag” to the word “stem”. The beauty of encryption is that it makes comparisons between words easier, since we don't need to deal with complex word changes. In our case we have used the “porter stemish” algorithm in both tweets and dictionary, whenever there is a need to compare.

6) *Exclusion Terminology:* Unique standalone words for general terms that do not contain additional information when used in a text and then declared invalid [19]. Examples include "a", "an", "the", "he", "is", "in", "always", etc. It is sometimes appropriate to remove these words because they do not contain additional information since they are used roughly in all sections of a text, for example when counting pre-tweet words from multiple occurrences and using this polarity to calculate general tweet sentiment over a set of words used in that tweet .
7) *Parts-Speech Tagging:* POS-Tagging is the process of tagging all words in a sentence, e.g. Is a noun a verb, an adjective, a adjective, a conjunction, a coordinating conjunction? etc.

D. Classification

Pattern classification is a process in which data is grouped into different classes according to specific patterns found in one class that vary to some extent by patterns found in other classes. The main purpose of our project is to design a classifier that accurately classifies tweets into the following four classes: constructive, negative, neutral, and confusing.
There are two types of overlap in this area: Sensitivity analysis and general analysis. Behavioral analysis addresses the separation of parts of a titter based on a given context, for example on a titter for "another 4 years of Australian grandmother and then moving to the USA: D" a desire to understand the situation that will identify Australia with negative emotions and USA in a good sense. On the other hand a general analysis of the subject meets the general understanding of the whole text (titter in this case). So in the aforementioned tweet as there is a perfect attitude, a normal group with common feelings can present it as positive. For our specific project we will be discussing the final case, i.e. the most common (standard) of analyzing the feelings of the whole tonso.
The classification method usually followed in this domain is a two-step process. The first Objectivity class is created that deals with classifying a tweet or phrase as a goal or less. After this we do Polarity Classization (in tweets only classified into individual categories) to find out if the titter is positive, negative or both (some researchers include both and some do not). This was introduced by Wilson et al. and reports improved accuracy rather than a simple one-step procedure [16].
We propose a novel approach that is slightly different than the method proposed by Wilson et al. [16]. We propose that in the first stage each tweet deals with two jump analysis: thececquer classifier and the polarity classifier. The first one can do it
There are two types of overlap in this area: Sensitivity analysis and general analysis. Behavioral analysis addresses the separation of parts of a titter based on a given context, for example on a titter for "another 4 years of Australian grandmother and then moving to the USA: D" a desire to understand the situation that will identify Australia with negative emotions and USA in a good sense. On the other hand a general analysis of the subject meets the general understanding of the whole text (titter in this case). So in the aforementioned tweet as there is a perfect attitude, a normal group with common feelings can present it as positive. For our specific project we will be discussing the final case, i.e. the most common (standard) of analyzing the feelings of the whole tonso.
The classification method usually followed in this domain is a two-step process. The first Objectivity class is created that deals with classifying a tweet or phrase as a goal or less. After this we do Polarity Classization (in tweets only classified into individual categories) to find out if the titter is positive, negative or both (some researchers include both and some do not). This was introduced by Wilson et al. and reports improved accuracy rather than a simple one-step procedure.
We propose a novel approach that is slightly different than the method proposed by Wilson et al. [16]. We propose that in the first stage each tweet deals with two jump analysis: thececquer classifier and the polarity classifier. The first one would try to differentiate the tweet between objective and consequential classes, while the latter would do that between the positive and negative categories. We use the shortlisted features of these programs as well.
So in step 2 we can treat each of these numbers as unique features in another program, where the feature size will be just 2. We use WEKA and use the following Machine Learning algorithms in this second framework to arrive at the best possible result:

1) K-K integration
2) Vector support machine
3) Postponement(LR)
4) Nearby Neighbors(K_N_N)
5) Naive Bayes
6) Govern Learning Support(R_B_C)

To better understand how this works we present a consensus test set from one of our validations in the 2-dimensional space mentioned above:

IV. CONCLUSION AND FUTURE RECOMMENDATIONS

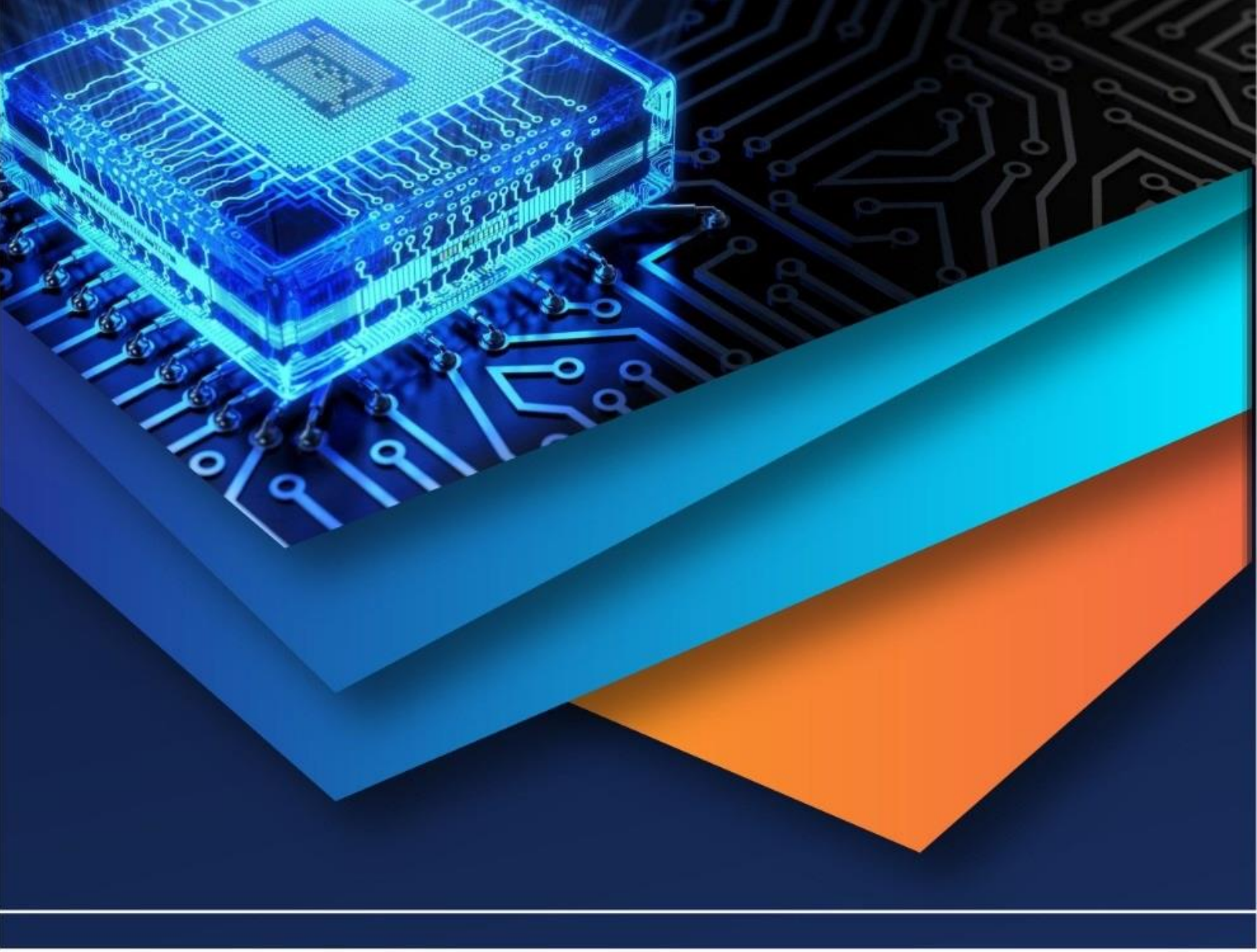
Major role of sentiment analysis, especially in micro-blogging domain, is in the developing phase. So we can propose a bunch of ideas may result in further improved performance.

We have demonstrate with unigram models; it also can be improved with the help of like closeness of the word with a negation type of word. We could mention a window prior to the word being analysis and the effect of negation may be highlighted if it follows within that window. The closer the negation word is to the unigram word whose prior polarity is to be find out, the more it should affect the polarity.

Now we are examining unigrams and effect of bigrams and trigrams may be carried out efficiently.

REFERENCES

[1] S. M. Metev and V. P. Veiko, Laser Assisted Microtechnology, 2nd ed., R. M. Osgood, Jr., Ed. Berlin, Germany: Springer-Verlag, 1998.
[2] J. Breckling, Ed., The Analysis of Directional Time Series: Applications to Wind Speed and Direction, ser. Lecture Notes in Statistics. Berlin, Germany: Springer, 1989, vol. 61.
[3] S. Zhang, C. Zhu, J. K. O. Sin, and P. K. T. Mok, "A novel ultrathin elevated channel low-temperature poly-Si TFT," IEEE Electron Device Lett., vol. 20, pp. 569–571, Nov. 1999.
[4] M. Wegmuller, J. P. von der Weid, P. Oberson, and N. Gisin, "High resolution fiber distributed measurements with coherent OFDR," in Proc. ECOC'00, 2000, paper 11.3.4, p. 109.
[5] R. E. Sorace, V. S. Reinhardt, and S. A. Vaughn, "High-speed digital-to-RF converter," U.S. Patent 5 668 842, Sept. 16, 1997.
[6] (2002) The IEEE website. [Online]. Available: http://www.ieee.org/
[7] M. Shell. (2002) IEEEtran homepage on CTAN. [Online]. Available: http://www.ctan.org/tex-archive/macros/latex/contrib/supported/IEEEtran/
[8] FLEXChip Signal Processor (MC68175/D), Motorola, 1996.
[9] "PDCA12-70 data sheet," Opto Speed SA, Mezzovico, Switzerland.
[10] A. Karnik, "Performance of TCP congestion control with rate feedback: TCP/ABR and rate adaptive TCP/IP," M. Eng. thesis, Indian Institute of Science, Bangalore, India, Jan. 1999.
[11] J. Padhye, V. Firoiu, and D. Towsley, "A stochastic model of TCP Reno congestion avoidance and control," Univ. of Massachusetts, Amherst, MA, CMPSCI Tech. Rep. 99-02, 1999.
[12] Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specification, IEEE Std. 802.11, 1997.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY