Inference Rating Framework for Sentiment Analysis Using SVM^{Light}

Tapan Biswas[1], Poonam Singh [2] and Binay Kumar Pandey [3]

[1]Tapan Biswas, Govindh Ballabh Pant University of Agriculure and Technology,Pantnagar, UK, INDIA
[1]tapan.biswas9997@yahoo.com
[2]Poonam Singh ,Govindh Ballabh Pant University of Agriculure and Technology,Pantnagar, UK, INDIA
[2]singh.poonam121@gmail.com
[3]Binay Kumar Pandey, Govindh Ballabh Pant University of Agriculure and Technology,Pantnagar, UK, INDIA
[3]binaydece@gmail.com

ABSTRACT

In this paper, we propose a noble approach for sentiment analysis with new perspective. Sentiment analysis/ Opinion mining is performed on credible data of a micro blogging site, Twitter. Twitter messages, tweets, are classified as positive, negative or neutral with respect to the parameters defined. We suggest, for extract credible and legitimate data, our path must be feasible, reliable and liable. We use a linear classifier machine learning algorithm, SVM. We also discuss the preprocessing steps needed in order to achieve high accuracy with reliability. The gist off this paper is to provide an approach which would enable us to unanimously conclude enormous corpus of tweets. Our methodology comprises of scripting computer language and a linear classifier machine learning algorithm.

Keywords— Sentiment Analysis, Opinion mining, Twitter Data Analysis, Machine Learning.

1. INTRODUCTION

What others think has always been an important piece of information. To know, how market is responding to a commodity, service, device, etc. or the people to bills, government policies, etc. on online could be quite important. To understand how important, we propose an approach which would enable us to unanimously conclude enormous corpus of data. But first, we need to understand the sentiments behind the text. Sentiment is the attitude, opinion or feeling toward something, such as a person, organization, product or location. Sentiment analysis (also known as opinion mining) refers to the use of natural language processing, text analysis and computational linguistics to identify and extract subjective information in source material.

Tweets (and micro blogs in general) are different from reviews primarily because of their purpose: while reviews represent summarized thoughts of authors, tweets are more casual and limited to 140 characters of text. Generally, users do not compose their tweets as thoughtfully as compared to reviews. Yet, they still offer companies an additional avenue to gather feedback. There has been some work by researchers in the area of phrase level and sentence level sentiment classification recently [1]. Previous research in sentiment analysis like Pang et al. [2] has analyzed the performance of different classifiers on movie reviews. Pang et al. also make use of a similar idea as ours, using star ratings as polarity signals in their training data. We show that we can produce comparable results on tweets with distant supervision and if domain permits, we can compare our results with existing platform values. Our practice emphasizes to retrieve only credible and legitimate data.

We have practiced inference rating approach to classify the tweets manually either as +1, 0 and -1. We have also used process of stemming as it reduces each word in the search index to its basic root or stem (e.g. 'blogging' to 'blog') so that variations on a word ('blogs', 'blogger', 'blogging', 'blog') are considered equivalent while creating a TDM (Term Document Matrix) for SVM^{Light} .The output, depending upon the domain,

is conclusive on its own or we can also use it to validate the verdicts obtained from existing platforms.

2. SENTIMENT ANALYSIS

Sentiment analysis (also known as opinion mining) refers to the use of natural language processing, text analysis and computational linguistics to identify and extract subjective information in source material. Sentiment analysis aims to determine the attitude of a speaker or a writer with respect to some topic or the overall contextual polarity of a document.

3. INFERENCE RATING

Inference is the act or process of deriving logical conclusions from premises known or assumed to be true or in layman's terms the process of inferring something. A tweet usually contains a mixture of positive and negative opinions towards different features of a product, a service or any object and inference rating aims at determining the overall sentiment implied by the user using the inferred knowledge gained by processing the tweet under consideration. There may be some predefined constraints that may be used so that the rating may gain some objective specific direction. Such task can be performed by aggregating the strengths of the opinion words in a review with respect to different sentiment classes, and then assigning an overall rating to the review to reflect the dominant sentiment class.

4. LINEAR CLASSIFIER

Statistical classification's goal is to use an object's characteristics to identify which class (or group) it belongs to. A linear classifier achieves this by making a classification decision based on the value of a linear combination of the characteristics. An object's characteristics are also known as feature values and are typically presented to the machine in a vector called a feature vector.

If the input feature vector to the classifier is a real vector $\vec{x}$, then the output score is:

$$y = f(\vec{w}.\vec{u}) = f\left(\sum_i w_i u_i\right) \dots\dots\dots\dots (1)$$

In Eq.1 $\vec{w}$ is a real vector of weights and *f* is a function that converts the dot product of the two vectors into the desired output. (In other words $\vec{w}$ is a one-form or linear functional mapping $\vec{x}$ onto R.) The weight vector $\vec{w}$ is learned from a set of labeled training samples. Often *f* is a simple function that maps all values above a certain threshold to the first class and all other values to the second class. A more complex *f* might give the probability that an item belongs to a certain class.

5. MACHINE LEARNING

Machine learning is a subfield of computer science that evolved from the study of pattern recognition and computational learning theory in artificial intelligence [3]. Machine learning explores the construction and study of algorithms that can learn from and make predictions on data. Machine learning can also be further distinguished as supervised machine learning, unsupervised machine learning and semi-supervised machine learning.

5.1. Supervised Machine Learning:

Supervised learning is the most common technique for training neural networks and decision trees. Both of these techniques are highly dependent on the information given by the pre-determined classifications. In the case of neural networks, the classification is used to determine the error of the network and then adjust the network to minimize it, and in decision trees, the classifications are used to determine what attributes provide the most information that can be used to solve the classification puzzle. We'll look at both of these in more detail, but for now, it should be sufficient to know that both of these examples thrive on having some "supervision" in the form of pre-determined classifications.

5.2. Unsupervised Machine Learning:

This approach proceeds by teaching the agent not by giving explicit categorizations, but by using some sort of reward system to indicate success. This type of training will generally fit into the decision problem framework because the goal is not to produce a classification but to make decisions that maximize rewards.

5.3. Semi-supervised Machine Learning:

Semi-supervised learning is a class of supervised learning tasks and techniques that also make use of unlabeled data for

training - typically a small amount of labeled data with a large amount of unlabeled data. Semi-supervised learning falls between unsupervised learning (without any labeled training data) and supervised learning (with completely labeled training data). Many machine-learning researchers have found that unlabeled data, when used in conjunction with a small amount of labeled data, can produce considerable improvement in learning accuracy. The acquisition of labeled data for a learning problem often requires a skilled human agent or a physical experiment. The cost associated with the labeling process thus may render a fully labeled training set infeasible, whereas acquisition of unlabeled data is relatively inexpensive.

6. APPLICATION PROGRAMMING INTERFACE

API, an abbreviation of application program interface, is a set of routines, protocols, and tools for building software applications. The API specifies how software components should interact and APIs are used when programming graphical user interface (GUI) components. A good API makes it easier to develop a program by providing all the building blocks. A programmer then puts the blocks together.

6.1. Types of APIs:

The API provided by the web services which also provide live feed of data from different source points categorize it as Representational State Transfer or REST API and Streaming API. We have used REST API.

7. APPROACH

We build a framework that treats feature extractors as two distinct components. We have used Twitter API to create an app to generate access tokens. Then we have used a scripting language Python and created code to fetch data from twitter database using prior generated access token. Domain of the data retrieved relies on the query term that we have we had passed in our code as parameter. There could be more than one query term for a single domain as some examples are shown in Table-1. We have used some modules like textmining, xlrd, etc. [4] in python library for processing the retrieved data.

TABLE-I

Domain	Query Term
Experience	#Yoga, @Yoga
Movie Review	#TheGreatGatsby, #TGGatsby, #Gatsby, @GreatGatsby, etc.
Product	#iPhone6, #iP6, @iphone6c

Table-1: Example of Query Term

We have to preprocess the retrieved data like we have to remove URL links, usernames (if present), re-tweet prefixes, etc. Retrieved tweets could be multi-lingual like English, Spanish, Portuguese, etc. But in our framework, it really doesn't matter as long as we could infer the correct sentiment behind that tweet. We have used Google Translate[5] to get the gist of the tweets which we were unable to comprehend.

We use SVM^{Light} [6] software with linear kernel for the classification purpose of our data set. We have chosen to feed a TDM (Term Document Matrix) as an input to the SVM^{Light}. A document-term matrix or term-document matrix is a mathematical matrix that describes the frequency of terms that occur in a collection of documents. In a document-term matrix, rows correspond to documents in the collection and columns correspond to terms [7]. The basic structure of a TDM is shown in Table-II.

TABLE-II

Entities	Attributes
<line>	<target> <term>:<value> ….#<info>
<target>	+1 | -1 | 0 | <float>
<term>	<integer>
<value>	<float>
<info>	<string>

Table-2: TDM structure

Ex: +1 60:1 128:1 338:1 362:1 374:1 406:2 465:1 499:1, where +1 is target value, 60, 128, 338, 362, 374, 406, 465 and 499 represents term value like 1, 2, etc. represents the frequency of the word in that tweet.

We have split the TDM into two. Comparative larger one is used to first train the machine or to make it learn. Once the machine is trained on the specified input data set, we can then finally make the machine to classify the data set made to be classified.

We have considered movie review domain. We have tried and succeeded in determining the sentiments behind the tweets on the above mentioned subject.

8. RESULTS

We have explored movie review domains. We have two kinds of outputs. First, the expected output that we have manually created and second that we have is from SVM^{Light} output. We have mapped corresponding values of tweets to make our results conclusive. The sentiments that we have assigned to the documents on each subject, was done on the basis of some predefined parameters. Some of them are listed in Table 3.

TABLE-III

Subject	Parameters (Tweets)	Value
Mad Max (Movie)	Litotes	+1
	Digressed	0
	Going to watch	0
	Mocking	-1
	Suggesting, etc.	+1

Table-3: Tweet Sentiment Parameters

We have plotted every input of expected output with respect to every input of corresponding SVM^{Light} output. And as we can see in Fig-1, the plot point co-ordinates are very discrete. Hence, we have used the linear regression method approach to understand the unanimity of the entire output range.

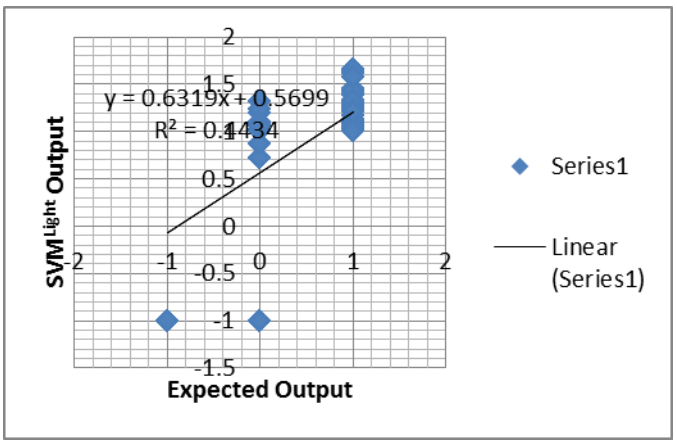

Fig-1: Mad Max Data Output Graph Plot

In statistics, linear regression is an approach for modeling the relationship between a scalar dependent variable y and one or more explanatory variables (or independent variable) denoted X. The case of one explanatory variable is called simple linear regression [8]. Fig-2[8] represents an example of a regression line for a given input.

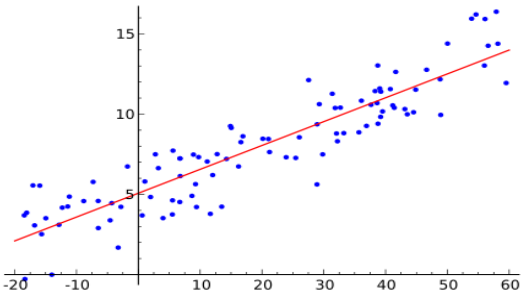

Fig-2: Example Figure for a Regression Line

Now, as we extend our work and explore our data range and monitor the average of both our output domains at equal intervals. SVMLight was trained with 1500 tweet documents. We have then monitored our expected output with the SVM^{Light}-classify prediction output. We have made SVM^{Light} to classify four different 500 tweet document for four different intervals. The averages at each interval are documented and are represented in Table-4.

TABLE-IV

Interval	Expected O/P Avg.	SVM^{Light} O/P Avg.
1	0.6621	0.96
2	0.652	0.954
3	0.66	0.98
4	0.6613	0.971

Table-4: Average Outputs

As we plot the graph for Table-4, the regression line passes through a point P with coordinate P(0.6588, 0.96625). Fig-3 represents all these coordinates on a graph plot. The coordinates of point P represents the mean value of both output averages.

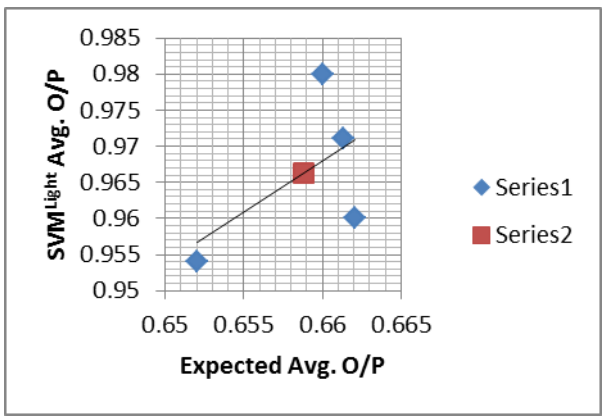

Fig-3: Mean Output

The P(0.6588, 0.96625) is represented by a square in Fig-3. This point P represents our ultimate goal of unanimously

concluding a value which represents the entire set of tweet documents.

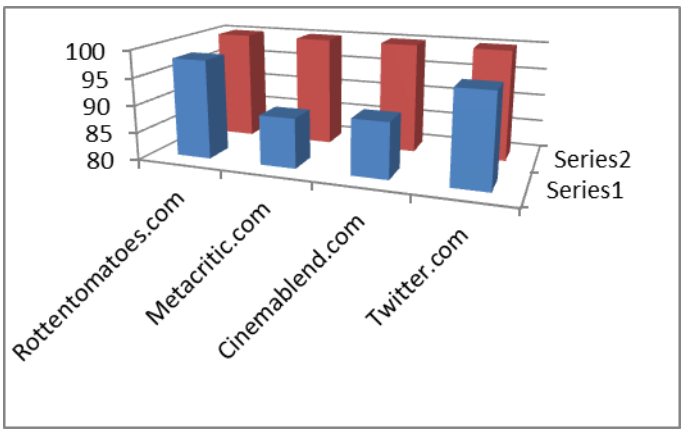

Fig-4: Mad Max Reviews

With the help of Fig-4, we can see the comparisons done between different movie reviewing websites and micro-blogging website Twitter.com.

9. CONCLUSIONS

We have shown that with our proposed framework and approach, we can unanimously conclude the large corpuses of Tweet documents of specific domain. We can now, with the help of proposed framework can also make viable comparison between unanimous review from the micro-blogging website Twitter.com and other existing movie reviewing websites.

10. ACKNOWLEDGMENT

We would like to express our sincere gratitude to the entire team of IJREST for providing us such a reputed journal. We express our best wishes for the Production team and other contributors for developing and maintaining the IJREST.

REFERENCES

[1] T. Wilson, J. Wiebe, and P. Ho®mann. Recognizing contextual polarity in phrase-level sentiment analysis. In Proceedings of Human Language Technologies Conference/Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP 2005), Vancouver, CA, 2005.

[2] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up? Sentiment classification using machine learning techniques. In Proceedings of the Conference on Empirical Methods in Natural Language Processing(EMNLP), pages 79{86, 2002}

[3] The Wikipedia homepage on Machine Learning. [Online]. Available: https://en.wikipedia.org/wiki/

[4] The Python.org homepage on xlrd [online]. Available: https://pypi.python.org/pypi/.

[5] The Translate Google website. [Online]. Available: https://translate.google.co.in/

[6] T. Joachims, Making large-Scale SVM Learning Practical. Advances in Kernel Methods - Support Vector Learning, B. Schölkopf and C. Burges and A. Smola (ed.), MIT-Press, 1999.

[7] The Wikipedia homepage on Document term matrix. [Online]. Available: https://en.wikipedia.org/wiki/

[8] The Wikipedia homepage on Linear regression. [Online]. Available: https://en.wikipedia.org/wiki/

