Twitter Sentimental Analysis based on Ordinal Regression

Rajashekar Nennuri¹, M Geetha Yadav², Y Sai Vahini³, Goda Sairam Prabhas⁴, V Rajashree⁵

¹,2,3,4,5Dept. of CSE, Institute of Aeronautical Engineering¹²³⁴⁵, Hyderabad
rajasekhar.nennuri@gmail.com¹ geethayadav22@gmail.com²
vahiniyammanuru@gmail.com³ Prabhasgoda123@gmail.com⁴
rajasreevytla15@gmail.com⁵

ABSTRACT

For associations and people with a profound social, political, or monetary interest in keeping up and fortifying their clout and notoriety, Twitter has become a goldmine. Sentiment analysis is the way toward characterizing and classifying the considerations and sentiments communicated in a source record. By performing this assessment investigation in a meticulous space, it is feasible to decide the force of area data on notion order. For feeling examination order, the proposed system utilizes the calculations Support Vector Regression (SVR), Decision Trees (DTs), and Random Forest (RF). The real execution of this structure depends on a twitter dataset unveiled by the NLTK corpora devices. The proposed approach will precisely identify ordinal relapse utilizing AI procedures. Keywords: Machine Learning, Twitter, sentiment analysis, ordinal regression.

1. Introduction

Around 200 million clients post 400 million tweets each day on Twitter, making it one of the biggest and most heterogeneous datasets of client created content. Micro blogging sites like Twitter have seen a massive expansion lately. As a result of this shift, firms and public partnership approach are increasingly trying to find a way to exploit Twitter for insights as to how people feel about their products or services. Symbolic procedure (otherwise called information base methodologies) and AI strategies are the two fundamental procedures utilized in sentimental analysis. An information base methodology requires a wide data set of predefined feelings just as a compelling data portrayal for characterizing sentiments. In an AI approach, a prepared data set is utilized to build a sentiment classifier that classifies opinions. In the investigation of Twitter sentiment, these methodologies are utilized. This methodology makes a feeling classifier that sorts feelings utilizing a training collection. We propose a strategy that incorporates preprocessing tweets, highlight extraction strategies, and the advancement of a scoring and balancing plan, trailed by the utilization of different AI procedures to order tweets into different classes.
2. Related Work
In paper [1] estimated public assessment from surveys with slant estimated from text. They broke down a few studies on buyer certainty and political assessment over the 2008 to 2009 period and they associated notion word frequencies in twitter messages. While the outcomes fluctuate across datasets, in a few cases the relationships as high as 80% and the outcomes feature the capability of text streams as a substitute and supplement for conventional surveying. [2] study expected to optimize N-gram based content component choice in sentiment analysis for business items in twitter through extremity dictionaries. This should be possible by consolidating reference based weighing with Naive Bayes classification of sentiments. This study is progressing yet results show potential. [3] distinguished feelings as a difficult issue. They introduced a framework that, given a theme, consequently discovered individuals who hold assessments about that subject and the sentiment of each opinion. The framework contains a module for deciding word assumption and another for joining opinions inside a sentence.

3. Procedure and Methodology
Twitter tweets about electronic goods are used to build a dataset. There are three stages to conducting a sentiment analysis. Preprocessing is completed in the first phase. Then, using related features, a feature vector is generated. Finally, tweets are categorized into positive and negative categories utilizing different classifiers. The conclusion is determined utilizing the quantity of tweets in each class.

![System Architecture](image)

**Figure 1: System Architecture**

3.1 Dataset Collection
The Twitter API is used to automatically capture tweets, which are then manually annotated as positive or negative. Using 600 positive and negative tweets, a dataset is generated.

|       | Positive | Negative | Total |
|-------|----------|----------|-------|
| Training | 500      | 500      | 1000  |
| Test    | 100      | 100      | 200   |

Table 1

3.2 Preprocessing Of Tweets
Tweets have already been pre-processed. Keyword extraction on Twitter is troublesome due to misspellings and dialect. A preprocessing step precedes feature extraction to prevent this. The preprocessing process
includes steps such as removing the URL and avoiding misspellings and dialect. To avoid misspellings, repeated characters are replaced with two examples.

| Text                                      | Label          |
|-------------------------------------------|----------------|
| 0  Follow fridday inte paris top engaged members co... | High Positive |
| 1  listen last night bleed amazing track scotland | High Positive |
| 2  congrats                                | moderate Positive |
| 3  everyone gon na talking abt rat boy today be | High Negative |
| 4  earth assume rain london likely influenced ove... | High Negative |
| 5  remember fab four 24 hour call damn miss much | High Negative |
| 6  thirsty                                  | Neutral        |

Table 2

3.3 Feature Extraction

Feature selection is not an easy job, and finding the most useful features for each domain necessitates a detailed investigation. The extraction of features is performed in two stages. The first move is to extract twitter-specific features. The related twitter features are hashtags and emoticons. A load of ‘1’ is assigned to optimistic emoticons, while a load of ‘-1’ is assigned to unconstructive emoticons. The twitter-specific features are extracted and then omitted from the tweets. Tweets are then treated as plain text, and the function vector is made up of eight related features. The eight features used are the part of speech (POS) tag, special keyword, negation instance, emoticon, number of optimistic keywords, number of unconstructive keywords, number of constructive hash tags, and number of unconstructive hash tags.

3.4 Balancing and Scoring Method

Implement a sentiment categorization model with the aim of evaluating tweet sentiments and classifying them as strong positive, slight positive, moderate, mild negative, or extreme negative, depending on the tweet's polarity. We add the values assigned to each component declared in the tweet to arrive at a general polarity score value for each tweet(t), which is intended to be used as a guideline.

3.5 Machine Learning Techniques

Numerous machine learning methods depend on supervised categorization ways, in which sentiment recognition is framed as a dual of positive and negative values. To train classifiers using this method, you'll need labeled info. Machine learning techniques are using a training set and a test set to perform classification. The dataset includes the input feature vectors and their equivalent class labels. This training
set is used to generate a classification model which attempts to categorize the input feature vectors into equivalent class labels. A test set is often used to check the replica by predicting the class labels of missing data selected features.

SVM (Support Vector Machine) is a powerful machine Learning algorithm both for regression and classification. The equivalence of the line in Support Vector Regression is \( y = wx + b \), which would be similar to Linear Regression. This unshakable line is referred to as hyper plane in SVR. Support Vectors are fact points on whichever side of the hyper plane that are neighboring to the hyper plane and can be used to plot the boundary line. Unlike most other regression models, that help to reduce the difference between the real and accepted value, the SVR strives to fit the best quote within such a threshold. As a consequence, we can tell that the SVR model focuses on ensuring the condition \(-a y -(wx+b)\).

![SVM Diagram](image)

Figure 2: SUPPORT VECTOR REGRESSION

The final decisions are estimated using just a decision tree that uses chart or model calls and a restricted executive speech. To reflect on anticipated, worst, and possible standards in a range of circumstances, the call tree assists in knowing, absorbing, and preparing for recent developments.

![Decision Tree](image)

Figure 3: DECISION TREE

Classification and regression is an rule for Random Forest. It is a label tree classifier set. It adjusts the over fitting to their coaching set thus the chance forest gives the benefit of over the call tree. Random forest gives an opinion in simplification of error and is resistant to over fitting.
4. Implementation and Results

Output Generation
Double, snap on run.bat file to run this project and obtain the underneath chart.

In the chart snap on ‘load NLTK Dataset’ to get the tweets dataset from NLTK library.

In the chart we are able to observe total 10000 tweets are present in the library, at present snap on ‘read NLTK tweets Data’ knob to study all tweets and to construct TFIDF vector. Ahead every knob snap you require to stay for some seconds to get desired result.
In the chart screen we can perceive total 8000 tweets vector are obtained for training reason and 2000 tweets are obtained for testing purpose. Now snap on ‘Run SVR Algorithm’ to construct train replica with the intention of dataset and to compute precision.

In the chart we are able to perceive SVR produced 0.71% guess precision, as of now snap on ‘Run Random Forest Algorithm’ key to compute its precision.
In the chart random forest got 0.88% precision, as of now snap on ‘run decision tree algorithm’ knob to compute its precision.

In the chart decision tree got 0.94% precision, now tick on ‘detect sentiment type’ knob and upload test tweets to forecast it sentiment. In test folder contained by test.txt you can observe there is no opinion label and claim will be noticed.
In the above screen uploading the test tweets case and underneath are the obtained values.

In the chart for every tweet, we be capable of observing the confidential sentiments. At present snap on ‘precision button’ to acquire underneath precision chart.
In the chart x-axis shows algorithm name and y-axis shows accuracy, as of above chart we can perceive decision tree get improved predictions contrast to others.

5. Conclusion and Future Scope
Customers' views on the crucial to success in the industry are analyzed using Twitter sentiment analysis. When it comes to classifying sentiment in tweets, machine learning techniques do a decent job. Symbolic approaches are more complex and time-consuming than Machine Learning techniques. These methods can be used to analyze Twitter sentiment. This research employs classifiers such as Support vector regression, Decision Trees, and Random Forest. The Twitter data set is used to optimize this process. The Decision Tree, on the other hand, has the highest accuracy, at 91.81 percent. We will strengthen this technique in the future by trying to use bigrams and trigrams. Furthermore, a variety of machine learning and deep learning methods, such as Deep Neural Networks, Convolutional Neural Networks, and regular Neural Networks, can be used to investigate this.

References
1. B. O’Connor, R. Balasubramanyan, B. R. Routledge, and N. A. Smith, “From tweets to polls: Linking text sentiment to public opinion time series,” in Proc. ICWSM, 2010, vol. 11, nos. 122–129, pp. 1–2.
2. M. A. Cabanlit and K. J. Espinosa, “Optimizing N-gram based text feature selection in sentiment analysis for commercial products in Twitter through polarity lexicons,” in Proc. 5th Int. Conf. Inf., Intell., Syst. Appl. (IISA), Jul. 2014, pp. 94–97.
3. S.-M. Kim and E. Hovy, “Determining the sentiment of opinions,” in Proc. 20th Int. Conf. Comput. Linguistics, Aug. 2004, p. 1367.
4. C. Whitelaw, N. Garg, and S. Argamon, “Using appraisal groups for sentiment analysis,” in Proc. 14th ACM Int. Conf. Inf. Knowl. Manage., Oct./Nov. 2005, pp. 625–631.
5. H. Saif, M. Fernández, Y. He, and H. Alani, “Evaluation datasets for Twitter sentiment analysis: A survey and a new dataset, the STS-Gold,” in Proc. 1st International Workshop Emotion Sentiment Social Expressive Media, Approaches Perspect. AI (ESSEM), Turin, Italy, Dec. 2013.

6. A. P. Jain and P. Dandannavar, “Application of machine learning techniques to sentiment analysis,” in Proc. 2nd Int. Conf. Appl. Theor. Comput. Commun. Technol. (iCATccT), Jul. 2016, pp. 628– 632.

7. A. Go, R. Bhayani, and L. Huang, “Twitter sentiment classification using distant supervision,” Processing, vol. 150, no. 12, pp. 1–6, 2009.

8. M. Bouazizi and T. Ohtsuki, “A pattern-based approach for multi-class sentiment analysis in Twitter,” IEEE Access, vol. 5, pp. 20617–20639, 2017.

9. MG Yadav, R Nennuri, “ Data Mining based Modern and Advanced Design and Develoement Applications”, International Journal of Pure and Applied Mathematics 119 (16), 4651-4658

10. R Nennuri, AK Chaitanya, LP Malyala, “ Implementation of data frame work system based on model driven architecture for MAS and Web based applications”, International Journal of Engineering & Technology 7 (2.20), 1-4

11. Rajasekhar Nennuri , Prathyusha Malyala , SaiTejaswi Thotakura,” Crime Prediction and Analysis Using Machine Learning”, International Journal of Advanced Science and Technology 29 (special issue), 9

12. YB M Geetha yadav Golla Swetha , Vasista Kumar, “Intrusion Detection Scheme Using Machine Learning”, iercsit-20

13. KS M Geetha Yadav , S Srihitha , C Tejaswi, “Clustering of Bigdata in Application Review Analysis”, International Journal of Advanced Science and Technology 29 (special issue)