International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Sentiment Classification System of Twitter Data for Positive and Negative Review Using Python

[1]R. Saranath, [2] Mr. K. Nirmal, M.C.A, M.Phil.,

[1,2] Assistant Professor, Master of Computer Application, Krishnasamy College of Engineering & Technology, Cuddalore.

ABSTRACT

Twitter is a popular social networking website where users posts and interact with messages known as “tweets”. This serves as a mean for individuals to express their thoughts or feelings about different subjects. Various different parties such as consumers and marketers have done sentiment analysis on such tweets to gather insights into products or to conduct market analysis. Furthermore, with the recent advancements in machine learning algorithms, the accuracy of our sentiment analysis predictions is able to improve. We will attempt to conduct sentiment analysis on “tweets” using various different machine learning algorithms. We also compare against an approach based on sentiment-bearing topic analysis, and find that semantic features produce better Recall and F score when classifying negative sentiment, and better Precision with lower Recall and F score in positive sentiment classification.

1. Introduction

The emergence of social media has given web users a venue for expressing and sharing their thoughts and opinions on all kinds of topics and events. Twitter, with nearly 600 million users1 and over 250 million messages per day,2 has quickly become a gold mine for organisations to monitor their reputation and brands by extracting and analysing the sentiment of the Tweets posted by the public about them, their markets, and competitors.

Sentiment analysis over Twitter data and other similar microblogs faces several new challenges due to the typical short length and irregular structure of such content. Two main research directions can be identified in the literature of sentiment analysis on microblogs. First direction is concerned with finding new methods to run such analysis, such as performing sentiment label propagation on Twitter follower graphs, and employing social relations for user-level sentiment analysis. The second direction is focused on identifying new sets of features to add to the trained model for sentiment identification, such as microblogging features including hash tags, emotions , the presence of intensifiers such as all-caps and character repetitions etc., and sentiment topic features.

Show an average of 4.78% increase in F score in comparison to using the commonPOS features alongside unigrams.Compare results with sentiment-bearing topic features and show that semanticfeatures improve F by 1.22% when identifying negative sentiment, but worsens Fby 2.21% when identifying positive sentiment.

2. Existing System

- We classify sentiments with the help of machine learning and natural language processing (NLP) algorithms, we use the datasets from Kaggle.
- The data provided comes with emoticons (emoji), usernames and hashtags which are required to be processed (so as to be readable) and converted into a standard form.
- We also need to extract useful features from the text such unigrams and bigrams which is a form of representation of the “tweet”.

Disadvantages

- Less number of Data set collection
- Feature handled is little complex
- Consuming huge time

3.Proposed System

- The proposed model of twitter data analysis will be implemented using Anaconda python.
- The tweets can be analysed and characterized based on the emotions used by the social users. We attempt to classify the polarity of the tweet where it is either positive or negative.
- The data provided comes with emoticons, usernames and hashtags which are required to be processed and converted into a standard form.

- It also needs to extract useful features from the text such unigrams and bigrams which is a form of representation of the “tweet”.

4.System Modules:

- Module 1: Dataset Collection
- Module 2: Pre-Processing
- Module 3: Implement the model
- Module 4: Sentiment Analysis
- Module 5: Evaluation

Module 1: Dataset Collection

A dataset (or data set) is a collection of data, usually presented in tabular form. Each column represents a particular variable. Each row corresponds to a given member of the dataset in question. It lists values for each of the variables, such as height and weight of an object. Each value is known as a datum. We have chosen to use a publicly-available Healthcare dataset which contains a relatively small number of inputs and cases. The data is arranged in such a way that will allow those trained in medical disciplines to easily draw parallels between familiar statistical and novel ML techniques. Additionally, the compact dataset enables short computational times on almost all modern computers**.**

Module 2: Pre-Processing

The Keras preprocessing package provides several common utility functions and transformer classes to change raw feature vectors into a representation that is more suitable for the downstream estimators.
In general, learning algorithms benefit from standardization of the data set. If some outliers are present in the set, robust scalers or transformers are more appropriate. The behaviors of the different scalers, transformers, and normalizers on a dataset containing marginal outliers is highlighted in compare the effect of different scalers on data with outliers.

1) Stopwords Removal:

A dictionary based approach is been utilized to remove stopwords from tweets. A generic stopword list containing 75 stopwords created using hybrid approach is used. The algorithm is implemented as below given steps. The target text is tokenized and individual words are stored in array. A single stop word is read from stopword list. The stop word is compared to target text in form of array using sequential search technique. If it matches , the word in array is removed , and the comparison is continued till length of array. After removal of stopword completely, another stopword is read from stopword list and again algorithm runs continuously until all the stopwords are compared. Resultant text devoid of stopwords is displayed, also required statistics like stopword removed, no. of stopwords removed from target text, total count of words in target text, count of words in resultant text, individual stop word count found in target text is displayed.

2) Stemming Technique:

After removing the unwanted words from the tweet, stemming technique is processed. Stemming is the process of reducing inflected (or sometimes derived) words to their word stem, base or root form generally a written word form. The stem need not be identical to the morphological root of the word; it is usually sufficient that related words map to the same stem, even if this stem is not in itself a valid root.

3) Tokeniztaion:

Tokenization is a step which splits longer strings of text into smaller pieces, or tokens. Larger chunks of text can be tokenized into sentences, sentences can be tokenized into words, etc. Further processing is generally performed after a piece of text has been appropriately tokenized. Tokenization is also referred to as text segmentation or lexical analysis. Sometimes segmentation is used to refer to the breakdown of a large chunk of text into pieces larger than words (e.g. paragraphs or sentences), while tokenization is reserved for the breakdown process which results exclusively in words.

Module 3: Implement the Model

The implementation model represents how a system (application, service, interface, etc.) works. It is often described with system diagrams and pseudocode to be later translated into real code. It is shaped by technical, organizations, and business constraints. Here we use sequential model, softmax. Sequence models the machine learning models that input or output sequences of data. Sequential data includes text streams, audio clips, video clips, time-series data and etc. The softmax function is often used in the final layer of a neural network-based classifier. Such networks are commonly trained under a log loss (or cross-entropy) regime, giving a non-linear variant of multinomial logistic regression. Softmax converts a real vector to a vector of categorical probabilities.

The elements of the output vector are in range (0, 1) and sum to 1.

Each vector is handled independently. The axis argument sets which axis of the input the function is applied along.

Softmax is often used as the activation for the last layer of a classification network because the result could be interpreted as a probability distribution.

The softmax of each vector x is computed as exp(x) / tf.reduce_sum(exp(x)).

The input values in are the log-odds of the resulting probability.

Arguments

- **x** : Input tensor.
- **axis**: Integer, axis along which the softmax normalization is applied.

Returns

Tensor, output of softmax transformation (all values are non-negative and sum to 1).

Raises

Value Error: In case dim(x) == 1.

Module 4: Sentiment Analysis

Sentiment analysis (or opinion mining) is a natural language processing technique used to determine whether data is positive, negative or neutral. Sentiment analysis is often performed on textual data to help businesses monitor brand and product sentiment in customer feedback, and understand customer needs.

This is usually referred to as fine-grained sentiment analysis, and could be used to interpret 5-star ratings in a review, for example:

- Very Positive = 5 stars
- Very Negative = 1 star

Emotion detection

This type of sentiment analysis aims to detect emotions, like happiness, frustration, anger, sadness, and so on. Many emotion detection systems use lexicons (i.e. lists of words and the emotions they convey) or complex machine learning algorithms.

MODEL 5: Evaluation

Model evaluation aims to estimate the generalization accuracy of a model on future (unseen/out-of-sample) data. In this model precision, support, accuracy and confusion matrix are used to evaluate the Sentiment analysis of Twitter.

Confusion matrix:

A confusion matrix is a table that is often used to **describe the performance of a classification model** (or "classifier") on a set of test data for which the true values are known.

5. Feasibility Study

The feasibility of the project is analyzed in this phase and business proposal is put forth N with a very general plan for the project and some cost estimates. During system analysis the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is not a burden to the company. For feasibility analysis, some understanding of the major requirements for the system is essential. Three key considerations involved in the feasibility analysis are

i. Economical Feasibility
ii. Technical Feasibility
iii. Social Feasibility

Economic Feasibility

This study is carried out to check the economic impact that the system will have on the organization. The amount of fund that the company can pour into the research and development of the system is limited. The expenditures must be justified. Thus, the developed system as well within the budget and this was achieved because most of the technologies used are freely available. Only the customized products had to be purchased.

Technical Feasibility

This study is carried out to check the technical feasibility, that is, the technical requirements of the system. Any system developed must not have a high demand on the available technical resources. This will lead to high demands on the available technical resources. This will lead to high demands being placed on the client. The developed system must have a modest requirement, as only minimal or null changes are required for implementing this system.

Social Feasibility

The aspect of study is to check the level of acceptance of the system by the user. This includes the process of training the user to use the system efficiently. The user must not feel threatened by the system, instead must accept it as a necessity. The level of acceptance by the users solely depends on the methods that are employed to educate the user about the system and to make him familiar with it. His level of confidence must be raised so that he is also able to make some constructive criticism, which is welcomed, as he is the final user of the system.

6. Testing

The purpose of testing is to discover errors. Testing is the process of trying to discover every conceivable fault or weakness in a work product. It provides a way to check the functionality of components, sub – assemblies, assemblies and/or a finished product It is the process of exercising software with the intent of ensuring that the Software system meets its requirements and user expectations and does not fail in an unacceptable manner. There are various types of test. Each test type addresses a specific testing requirement.

Unit Testing

Unit testing involves the design of test cases that validate that the internal program logic is functioning properly, and that program inputs produce valid outputs. All decision branches and internal code flow should be validated. It is the testing of individual software units of the application .it is done after the completion of an individual unit before integration. This is a structural testing, that relies on knowledge of its construction and is invasive. Unit tests perform basic tests at component level and test a specific business process, application, and/or system configuration. Unit tests ensure that each unique path of a business process performs accurately to the documented specifications and contains clearly defined inputs and expected results.

Integration Testing

Integration tests are designed to test integrated software components to determine if they actually run as one program. Testing is event driven and is more concerned with the basic outcome of screens or fields. Integration tests demonstrate that although the components were individually satisfaction, as shown by successfully unit testing, the combination of components is correct and consistent. Integration testing is specifically aimed at exposing the problems that arise from the combination of components.

7. Conclusion and Future Work

The analysis of Twitter data is being done in different points of view to mine the opinion or sentiment. Our proposed approach classify the tweets as Positive and Negative tweets which further helps in sentiment analysis and uses that sentiment analysis for further decision making. For our prototype, Twitter API is utilized to gather data in real-time. The prototype back-end tests on retrieving and processing the API data indicate that it is successful in gathering large amounts of data from popular search terms in real-time. We will use various machine learning algorithms to conduct sentiment analysis using the extracted features. However, just relying on individual models did not give a high accuracy so we pick the top few models to generate a model.

Reference

1. Joshi, Shaunak, and Deepali Deshpande. "Twitter sentiment analysis system." arXiv preprint arXiv:1807.07752 (2018).
2. Bhuta, Sagar, Avit Doshi, Uehit Doshi, and Meera Narvekar. "A review of techniques for sentiment analysis Of Twitter data." In 2014 International conference on issues and challenges in intelligent computing techniques (ICICT), pp. 583-591. IEEE, 2014.
3. Sailunaz, Kashfia, and Reda Alhajj. "Emotion and sentiment analysis from Twitter text." Journal of Computational Science 36 (2019): 101003.
4. Sarlan, Aliza, Chayanit Nadam, and Shuib Basri. "Twitter sentiment analysis." In Proceedings of the 6th International conference on Information Technology and Multimedia, pp. 212-216. IEEE, 2014.
5. Angiani, Giulio, Laura Ferrari, Tomaso Fontanini, Paolo Fornacciari, Eleonora Iotti, Federico Magliani, and Stefano Manicardi. "A Comparison between Preprocessing Techniques for Sentiment Analysis in Twitter." In KDWeb. 2016.
6. Bouazizi, Mondher, and Tomoaki Ohtsuki. "Sentiment analysis: From binary to multi-class classification: A pattern-based approach for multi-class sentiment analysis in Twitter." In 2016 IEEE International Conference on Communications (ICC), pp. 1-6. IEEE, 2016.
7. Hu, Xia, Jiliang Tang, Huiji Gao, and Huan Liu. "Unsupervised sentiment analysis with emotional signals." In Proceedings of the 22nd international conference on World Wide Web, pp. 607-618. 2013.
8. Goel, Ankur, Jyoti Gautam, and Sitesh Kumar. "Real time sentiment analysis of tweets using Naive Bayes." In 2016 2nd International Conference on Next Generation Computing Technologies (NGCT), pp. 257-261. IEEE, 2016.
9. Kumar, Akshi, and Arunima Jaiswal. "Systematic literature review of sentiment analysis on Twitter using soft computing techniques." Concurrency and Computation: Practice and Experience 32, no. 1 (2020): e5107.
10. Bouazizi, Mondher, and Tomoaki Ohtsuki. "A pattern-based approach for multi-class sentiment analysis in Twitter." IEEE Access 5 (2017): 20617-20639.
11. Pandey, Avinash Chandra, Dharmveer Singh Rajpoot, and Mukesh Saraswat. "Twitter sentiment analysis using hybrid cuckoo search method." Information Processing & Management 53, no. 4 (2017): 764-779.
12. Pandarachalil, Rafeeque, Selvaraju Sendhilkumar, and G. S. Mahalakshmi. "Twitter sentiment analysis for large-scale data: an unsupervised approach." Cognitive computation 7, no. 2 (2015): 254-262.
13. Kumar, Akshi, and Geetanjali Garg. "Sentiment analysis of multimodal twitter data." Multimedia Tools and Applications 78, no. 17 (2019): 24103-24119.
14. Bao, Yanwei, Changqin Quan, Lijuan Wang, and Fuji Ren. "The role of pre-processing in twitter sentiment analysis." In International conference on intelligent computing, pp. 615-624. Springer, Cham, 2014.