International Journal of Research Publication and Reviews

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

Automated Real-Time Twitter Sentiment Analysis Web App

[1]G. Priya, [2] Dr. J. Sreerambabu, [3]D. Rajkumar

[1]PG Scholar,[2] Head of the Department, [3] Assistant Professor
Master of Computer Applications Department, Thanthai Periyar Government Institute of Technology, Vellore-2.

ABSTRACT:

With the evolving conduct of various kinds of social networking web sites like Instagram, twitter, snapchat etc , the statistics published with the aid of using human beings i.e the customers of a specific social web website online is growing drastically . So plenty in order that nearly hundreds of thousands and billions of statistics might also additionally it's textual, video or audio is published in step with day. This is due to the fact there are hundreds of thousands of customers of a specific web website online. These customers intend to percentage their thoughts, perspectives associated with any subject matter in their choosing.

Some of those customers even put up in vain. These posts are quick for this reason simplest intended to specific a specific view of a specific consumer concerning a specific thing. In this paper we purpose to derive the emotions at the back of those posts. For this we've selected twitter as a social networking web website online. The posts on this social networking web website online are called tweets. In this paper we scrutinize strategies of preprocessing and extraction of twitter statistics the usage of python after which teach in addition to take a look at this statistics towards a classifier that allows you to derive the emotions at the back of tweets.

Key word: Sentiment Analysis, Tweets, Train & Test data, Preprocessing, Data set.

Introduction

Microblogging web sites, in these days's international have come to be a sea of facts for analysts to prey on. This is due to the fact maximum of the people these days are linked to a few type of microblogging web website online in which they pull out all of the hype they sense concerning anything. It won't be incorrect to mention that during a few manner those Microblogging webweb sites have given a proper to speech to each man or woman who can get admission to them. People from various elements of the sector freely discuss , comment , submit their reviews approximately any subject matter in their selecting in actual time .These blogs are broadly speaking a bitch expressing a poor vibe Or an appreciation expressing a advantageous vibe towards any subject matter in their selecting . The subjects humans submit approximately might be a product from an employer inclusive of a computer or a phone. Or it is able to be a well-known entity Or every other thing. Most of the main companies in these days's technology have hired analysts who've a task to derive feelings of humans in the back of those posts. This facilitates them to get a right assessment About their product or enterprise which facilitates them understand public call for and the changes they Need to make a good way to make higher product in future. Therefore from the dialogue above it is able to be concluded that those micro-running a blog web sites should come to be an asset to extraordinary companies public or non-public if evaluation of sentiment might be applied on them.

Sentiment evaluation additionally called evaluation of emotions is an beneficial device for reading extraordinary web sites in which humans submit their reviews concerning a subject of hobby .With the assist of this type of evaluation companies can reap the feelings of the humans which they submit as tweets or as feedback or while assessment concerning a selected entity or fabricated from hobby to them .This is going according with[10] who says , nearly 87% humans having a reference to net take a look at critiques earlier than purchase. This approach might be used for extraordinary functions inclusive of politicians should use it for reading what type of sentiments humans from extraordinary regions are sporting toward him/her and consequently should make investments extra in the ones regions. An instance of that is current Trump elections, in which he employed a set of analysts for this precise purpose. Sentiment evaluation can also be implemented with inside the discipline of commercial enterprise marketing. With the assist of this era extraordinary commercial enterprise companies seize the emotions of humans concerning their merchandise and of that in their competitors. Organizations rent there techniques with accordance to this know-how only. Leaving marketplace studies aside, evaluation of sentiments should play a critical element in Service industries As it is able to examine a complete fledged patron enjoy and will display patron feeling, that can show to be very beneficial.

PURPOSE OF THE SYSTEM

The accuracy is high. Time saving. Instead of storing the facts previous to the evaluation, it could get actual time facts from twitter via way of means of giving a hashtag or username to investigate the tweets of someone or a detailed hashtag.

2. SYSTEM ANALYSIS

EXISTING SYSTEM

The existing system „Sentiment Analysis" takes the static data which is already extracted from a social media platform. The data extracted is stored in a csv file or Excel file which is the input to the program or application. For each statement the program analyses, the output would be a floating-point number which is termed as polarity. The polarity values range from -1 to +1. Based on the polarity obtained the program determines the emotion of the statement. The emotion is classified as positive, negative, neutral.

If polarity>0 then the emotion is positive.

If polarity= 0 then the emotion is neutral.

If polarity<0 then the emotion is negative.

PROPOSED SYSTEM

This gadget offers with appearing features dynamically thru an internet social media i.e., Twitter. Twitter posts of digital merchandise creates a dataset. Tweets are brief messages with slang phrases and misspellings. So, the sentence degree sentiment evaluation is accomplished. This may be accomplished in seven phases. In the primary phase, enter records is given. Here the enter records refers to a username or a hashtag. Then, the quantity of tweets to be analyzed are specified. Those tweets are retrieved from the twitter database. Then with inside the third phase, the retrieved twitter records is saved in a database. In fourth phase, the tweet is processed. This step is accomplished earlier than characteristic extraction. Processing steps encompass disposing of URLs, disposing of stop-phrases, averting mis-spellings and slang phrases. Mis-spellings rectangular degree averted through commutation chronic characters with 2 occurrences. Slang phrases make a contribution considerable to the sensation of a tweet. Hence, a slang phrase lexicon is maintained to replace slang phrases taking place in tweets with their related meanings. Next element is characteristic extraction. A characteristic vector is shaped mistreatment applicable options.

3. DEVELOPMENT ENVIRONMENT

HARDWARE REQUIREMENTS

RAM : eight GB Ram

Processor : Intel i5 Processor or More

Hard Disk : 1TB

GUI: 2GB

SOFTWARE REQUIREMENTS

Front End : JavaScript

Operating gadget : Windows10

Platform : Anaconda Navigator

Backend : Machine Learning

Framework : Textblob & Tweepy, Twitter API, Flask

4. MODULE DESCRIPTION

DATA COLLECTION

Data collection is the first phase for analysis as there needs to be data for us to do analysis on. In our experimentations we have used python programming language as a tool. Being that said, data collection in this particular analysis could be carried out in two ways. First way is to collect preorganized data from different sites such as kraggle . On these sites this preorganized data is uploaded by the developers of sites themselves or is posted by different researchers for free . All one desires to do to collect this records is to create a loose account on those sites. Second way is to manually extract data from twitter using some API available for twitter. For this we have chosen tweepy as an API for extraction of tweets. Tweepy does not compatible with the new versions of python (python 3.7) . So for using this particular API an older version of python is needed(python 2.7). To access tweets on twitter using API first we need to authenticate the console from which we are trying to access twitter.

DATA PREPROCESSING

The pre-processing of data implies the processing of raw data into a more convenient format which could be fed to a classifier in order to better the accuracy of the classifier. Here, in our case the raw data which is being extracted from twitter using an API is initially totally unstructured and bogus as the availability of various useless characters seems very common in it. For this matter we remove all the unnecessary characters and words from this data using a module in python known as Regular Expressions, are for short. This module adopts symbolic techniques to represent different noise in the data and therefore makes it easy to drop them. Specifically in twitter terminology there are various common useless phrases and spelling mistakes present in the data, which need to be removed to boost the accuracy of our resultant. These could be summoned up as follows:

- Hash tags: these are very common in tweets. Hash tags represent a topic of interest about which the tweet is being written. Hashtags look something like #topic.
- @Usernames: these represent the user mentions in a tweet. Some times a tweet is written and then is associated with some twitter user, for this purpose these are used.
- Retweets(RT): as the name suggests retweets are used when a tweet is posted twice by same or different user.
- Emoticons: these are very commonly found in the tweets. Using punctuations facial expressions are formed in order to represent the a smile or other expressions, these are known as emoticons.
- Stop words: stop words are those word which are useless when it comes to sentiment analysis. Words such as it, is, the etc are known as stop words.

FEATURE SELECTION

This work have used different features for the classification of the tweets, in our experimentations similar feature are being used. These features include Unigram, Bigram, N-gram, POS tagging, Subjective, objective features and so on. NLTK short for Natural Language Tool Kit is another module available in python which also open source and could be used for extraction of these features.

MODEL SELECTION

Once the data is being pre-processed, this data is to be fed to a classification model for further processing. There are different classification algorithms on which these models are built on. In this paper, we have chosen k-nearest neighbour model to perform the classification. KNN or k-Nearest Neighbour algorithm represents a machine learning technique used for classifying a set of data into its given target values (in our case positive , neutral or negative).KNN could also be used for regression problems but is widely used for classification problems. Now, any classification model needs a target set on which we train the model for its further use. As for mentions in the literature survey section most of them have manually set these target values to positive, negative or null. For this paper we have used a library in python known as Textblob to automatically set the target for each tweet. The data set then is divided into two halfs training set and testing set. The data set used by us in our experimentations consisted of 2928 tweets so we segregated it into training and testing data. Training portion consisted of 2343 tweets whereas the test set consisted of 585 tweets. Now this training as well as test set needs to be transformed into binary values so as to be fed to the model. The models don't understand any values other than the binary. For this we have used another module of the python known as sklearn which contains many classification model as well as different encoders in it. For this paper this library is being for model selection, label encoding as well as model evaluation which would be mentioned in next section.

MODEL EVALUATION

One of the most common and appropriate technique used for evaluation of a classifier is through confusion matrix. By applying this technique we can derive the generalized evaluation parameters. These parameters include:

Accuracy: Accuracy of a classifier indicates how accurately the classifier has predicted the result.

Precision: Precision shows how often the result that is being predicted by the classifier, when it indicates true is correct.

Recall: It indicates the true positive rate of the classifier.

F1 score: It indicates the weighed average of recall and precision.

5. SYSTEM ARCHITECTURE

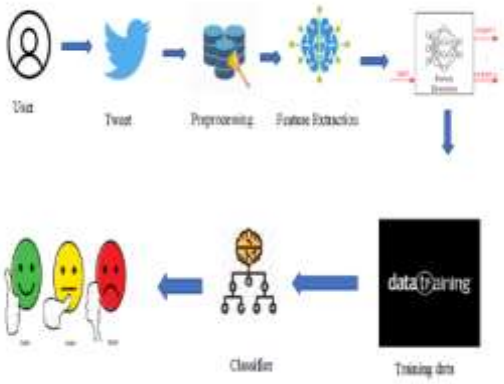

Fig: System Architecture

DATA FLOW DIAGRAM

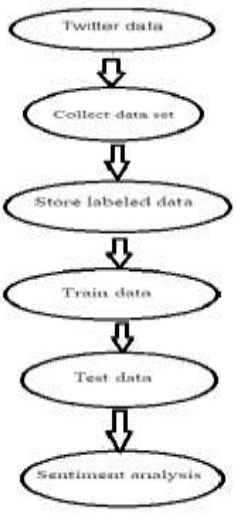

Fig: Flow System

6. CONCLUSION:

We tried to build a sentiment analysis system by studying and implementing algorithms of machine learning. We implemented Naive Bayes and Maximum Entropy algorithms. Baseline model performed the worst with no doubt as it had least number of features. The modular system we've built can easy be scaled for new algorithms be it in Machine Learning, Deep learning or Natural Language Processing. Sentiment analysis system is an active field of research and we can still further improve our system by working more on the algorithms, trying out different things in preprocessing and checking which ones get the best precision metrics.

7. FUTURE ENHANCEMENT:

Handling Emotion Ranges: We can enhance and educate our fashions to deal with a number of sentiments. Tweets don't continually have superb or poor sentiment. At instances they will don't have any sentiment i.e. neutral. Sentiment also can have gradations just like the sentence, This is good, is superb however the sentence, This is extraordinary. Is really greater superb than the first. We can consequently classify the sentiment in ranges; say from -2 to +2. 6.1.2

Using symbols: During our pre-processing, we discard maximum of the symbols like commas, full-stops, and exclamation mark. These symbols can be beneficial in assigning sentiment to a sentence.

References

[1]. 1995-1997: LSTM was proposed by Sepp Hochreiter and Jürgen Schmidhuber. By introducing Constant Error Carousel (CEC) units, LSTM deals with the vanishing gradient problem. The initial version of LSTM block included cells, input and output gates.1999: Felix Gers and his advisor Jürgen Schmidhuber and Fred Cummins introduced the forget gate (also called “keep gate”) into LSTM architecture, enabling the LSTM to reset its own state.

[2]. 2000: Gers & Schmidhuber & Cummins added peephole connections (connections from the cell to the gates) into the architecture. Additionally, the output activation function was omitted.

[3]. 2009: An LSTM based model won the ICDAR connected handwriting recognition competition. Three such models were submitted by a team led by Alex Graves. One was the most accurate

[4]. model in the competition and another was the fastest.

[5]. 2013: LSTM networks were a major component of a network that achieved a record 17.7% phoneme error rate on the classic TIMIT natural speech dataset.

[6]. 2014: Kyunghyun Cho et al. put forward a simplified variant called Gated recurrent unit (GRU).

[7]. 2015: Google started using an LSTM for speech recognition on Google Voice. According to the official blog post, the new model cut transcription errors by 49%.

[8]. 2016: Google started using an LSTM to suggest messages in the Allo conversation app. In the same year, Google released the Google Neural Machine Translation system for Google Translate which used LSTMs to reduce translation errors by 60%. Apple announced in its Worldwide Developers Conference that it would start using the LSTM for quicktype in the iPhone and for Siri. Amazon released Polly, which generates the voices behind Alexa, using a bidirectional LSTM for the text-to-speech technology

[9]. 2017: Facebook performed some 4.5 billion automatic translations every day using long shortterm memory networks. Researchers from Michigan State University, IBM Research, and Cornell University published a study in the Knowledge Discovery and Data Mining (KDD) conference. Their study describes a novel neural network that performs better on certain data sets than the widely used long short-term memory neural network. Microsoft reported reaching 94.9% recognition accuracy on the Switchboard corpus, incorporating a vocabulary of 165,000 words. The approach used "dialog session-based long-short-term memory".

[10]. 2019: Researchers from the University of Waterloo proposed a related RNN architecture which represents continuous windows of time. It was derived using the Legendre polynomials and outperforms the LSTM on some memory-related benchmarks. An LSTM model climbed to third place on the in Large Text Compression Benchmark.A RNN using LSTM units can be trained in a supervised fashion, on a set of training sequences, using an optimization algorithm, like gradient descent, combined with backpropagation through time to compute the gradients needed during the optimization process, in order to change each weight of the LSTM network in proportion to the derivative of the error (at the output layer of the LSTM network) with respect to corresponding weight. A problem with using gradient descent for standard RNNs is that error gradients vanish exponentially quickly with the size of the time lag between important events. This is due to {\displaystyle \lim _{n\to \infty }W^{n}=0}{\displaystyle \lim _{n\to \infty }W^{n}=0} if the spectral radius of {\displaystyleW}W is smaller than 1. However, with LSTM units, when error values are back-propagated from the output layer, the error remains in the LSTM unit's cell. This "error carousel" continuously feeds error back to each of the LSTM unit's gates, until they learn to cut off the value.