Sentiment Classification system of Twitter Data for Positive and Ukraine sentimental analysis project Negative Reviews using Python.

Gosala Saileela

Department of Computer Science, Sri Venkateswara University, Tirupati

***Abstract*—** *Nowadays, people from all around the world use social media sites to share information. Twitter for example is a platform in which users send, read posts known as 'tweets' and interact with different communities. Users share their daily lives, post their opinions on everything such as brands and places. Python is simple yet powerful, high-level, interpreted and dynamic programming language, which is well known for its functionality of processing natural language data by using NLTK (Natural Language Toolkit). NLTK is a library of python, which provides a base for building programs and classification of data. NLTK also provide graphical demonstration for representing various results or trends and it also provide sample data to train and test various classifier respectively. Companies can benefit from this massive platform by collecting data related to opinions on them. The main aim of this project is to classify the sentimental analysis for Ukraine sentimental analysis project using machine learning techniques. This Globally people are using social media platforms to share their ideas and information related to different topics every day. Twitter is one of the most popular social media platforms to send and read posts to communicate with others known as "tweets". Peoples share their ideas, reviews, experiences, and post their opinions on a particular topic or issue. This project aims to build a model that performs a sentimental analysis of people's opinions related to the social issues of Ukraine war. A dataset of tweets has been collected from Twitter by using a twitter scraper in python programming and then cleaned the dataset using the nltk library to remove noise from the dataset to show the result of the system.*

I. INTRODUCTION

Social Media like Instagram, Facebook, Twitter, WhatsApp, Telegram are the most powerful platforms that allow users to communicate with all over the world. Users can post their opinions and ideas about current issues, product reviews, share their moments, even share social issues. People post their opinions related to these issues through different social media platforms. Twitter is one of the most famous platforms for users to communicate with people. Sentiment analysis, for classifying specific words into positive or negative and neutral. For Ukraine sentimental analysis project using machine learning techniques.

1.1 Artificial Intelligence:

Artificial intelligence (AI) is the ability of a computer program or a machine to think and learn. It is also a field of study which tries to make computers "smart". As machines become increasingly capable, mental facilities once thought to require intelligence are removed from the definition. AI is an area of computer sciences that emphasizes the creation of intelligent machines that work and reacts like humans. Some of the activities computers with artificial intelligence are designed for include: Face recognition, Learning, Planning, Decision making etc.,

Artificial intelligence is the use of computer science programming to imitate human thought and action by analysing data and surroundings, solving or anticipating problems and learning or self-teaching to adapt to a variety of tasks.

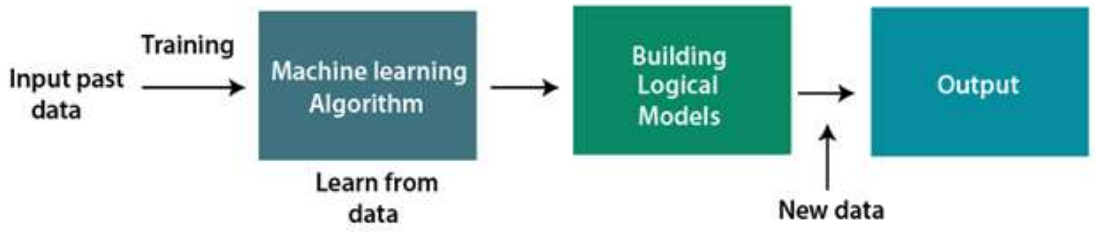
1.2 Machine Learning

Machine learning is a growing technology which enables computers to learn automatically from past data. Machine learning uses various algorithms for building mathematical models and making predictions using historical data or information. Currently, it is being used for various tasks such as image recognition, speech recognition, email filtering, Facebook auto-tagging, recommender system, and many more.

Machine Learning is said as a subset of artificial intelligence that is mainly concerned with the development of algorithms which allow a computer to learn from the data and past experiences on their own. The term machine learning was first introduced by Arthur Samuel in 1959. We can define it in a summarized way as: "Machine learning enables a machine to automatically learn from data, improve performance from experiences, and predict things without being explicitly programmed".

A Machine Learning system learns from historical data, builds the prediction models, and whenever it receives new data, predicts the output for it. The accuracy of predicted output depends upon the amount of data, as the huge amount of data helps to build a better model which predicts the output more accurately.

Suppose we have a complex problem, where we need to perform some predictions, so instead of writing a code for it, we just need to feed the data to generic algorithms, and with the help of these algorithms, machine builds the logic as per the data and predict the output. Machine learning has changed our way of thinking about the problem. The below block diagram explains the working of Machine Learning algorithm:



1.3 Machine learning

1. Machine learning is an application of artificial intelligence that involves algorithms and data that automatically analyze and make decision by itself without human intervention. It describes how computer perform tasks on their own by previous experiences. Therefore, we can say in machine language artificial intelligence is generated on the basis of experience.
2. The difference between normal computer software and machine learning is that a human developer hasn't given codes that instructs the system how to react to situation, instead it is being trained by a large amount of data.

1.4 Types of Machine Learning

- Supervised learning
- Unsupervised learning
- Reinforcement learning

- **Supervised learning**

Supervised learning is a technique where the program is given labelled input data and the expected output data. It gets the data from training data containing sets of examples.

They generate two kinds of results,

Classification: They notify the class of the data it is presented with**.**

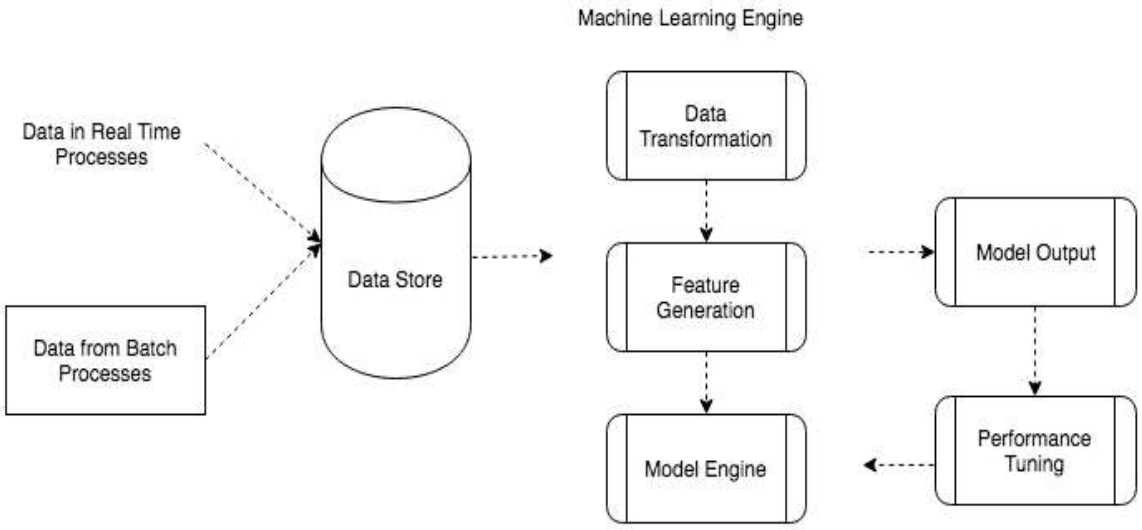
Regression: They expect the product to produce a numerical value.

UNSUPERVISED LEARNING

This type of algorithm consists of input data without labelled response. There will not be any preexisting labels and human intervention is also less. It is mostly used in exploratory analysis as it can automatically identify the structure in data.

REINFORCEMENT LEARNING

This model is used in making a sequence of decisions. It is a learning by interacting with the environment. It is based on the observation that intelligent agents tend to repeat the action that are rewarded for and refrain from action that are punished for. It can be said that it is a trial-and-error method in finding the best outcome based on experience.



II. LITERATURE REVIEW

[1] Title: Brand-Related Twitter Sentiment Analysis Using Feature Engineering and the Dynamic Architecture for Artificial Neural Networks

Authors: David Zimbra; M. Ghiassi; Sean Lee

Description: We present an approach to brand-related Twitter sentiment analysis using feature engineering and the Dynamic Architecture for Artificial Neural Networks (DAN2). The approach addresses challenges associated with the unique characteristics of the Twitter language, and the recall of mild sentiment expressions that are of interest to brand management practitioners. We demonstrate the effectiveness of the approach on a Starbucks brand-related Twitter data set. The feature engineering produced a final tweet feature representation consisting of only seven dimensions, with greater feature density. Two sets of experiments were conducted in three-class and five-class tweet sentiment classification. We compare the proposed approach to the performances of two state-of-the-art Twitter sentiment analysis systems from the academic and commercial domains. The results indicate that the approach outperforms these state-of-the-art systems in both three-class and five-class tweet sentiment classification by wide margins, with classification accuracies above 80% and excellent recall of mild sentiment tweets.

[2] Title: Targeted Twitter Sentiment Analysis for Brands Using Supervised Feature Engineering and the Dynamic Architecture for Artificial Neural Networks

Authors: Manoochehr Ghiassi

Description: Social media communications offer valuable feedback to firms about their brands. We present a targeted approach to Twitter sentiment analysis for brands using supervised feature engineering and the dynamic architecture for artificial neural networks. The proposed approach addresses challenges associated with the unique characteristics of the Twitter language and brand-related tweet sentiment class distribution. We demonstrate its effectiveness on Twitter data sets related to two distinctive brands. The supervised feature engineering for brands offers final tweet feature representations of only seven dimensions with greater feature density. Reducing the dimensionality of the representations reduces the complexity of the classification problem and feature sparsity. Two sets of experiments are conducted for each brand in three-class and five-class tweet sentiment classification. We examine five-class classification to target the mild sentiment expressions that are of particular interest to firms and brand management practitioners. We compare the proposed approach to the performances of two state-of-the-art Twitter sentiment analysis systems from the academic and commercial domains. The results indicate that it outperforms these state-of-the-art systems by wide margins, with classification F_1-measures as high as 88 percent and excellent recall of tweets expressing mild sentiments. Furthermore, they demonstrate the tweet feature representations, though consisting of only seven dimensions, are highly effective in capturing indicators of Twitter sentiment expression. The proposed approach and vast majority of features identified through supervised feature engineering are applicable across brands, allowing researchers and brand management practitioners to quickly generate highly effective tweet feature representations for Twitter sentiment analysis on other brands.

[3] TITLE: Twitter brand sentiment analysis: A hybrid system using *n*-gram analysis and dynamic artificial neural network

Authors: M.GhiassiaJ.SkinnerbD.Zimbraa

Description: Twitter messages are increasingly used to determine consumer sentiment towards a brand. The existing literature on Twitter sentiment analysis uses various feature sets and methods, many of which are adapted from more traditional text classification problems. In this research, we introduce an approach to supervised feature reduction using *n*-grams and statistical analysis to develop a Twitter-specific lexicon for sentiment analysis. We augment this reduced Twitter-specific lexicon with brand-specific terms for brand-related tweets. We show that the reduced lexicon set, while significantly smaller (only 187 features), reduces modeling complexity, maintains a high degree of coverage over our Twitter corpus, and yields improved sentiment classification accuracy. To demonstrate the effectiveness of the devised Twitter-specific lexicon compared to a traditional sentiment lexicon, we develop comparable sentiment classification models using SVM. We show that the Twitter-specific lexicon is significantly more effective in terms of classification recall and accuracy metrics. We then develop sentiment classification models using the Twitter-specific lexicon and the DAN2 machine learning approach, which has demonstrated success in other text classification problems. We show that DAN2 produces more accurate sentiment classification results than SVM while using the same Twitter-specific lexicon.

[4] TITLE: Real-time Twitter Sentiment Analysis using 3-way classifier

Authors: Alaa S. Al Shammari

Description: Sentiment analysis or opinion mining is a critical issue where a huge amount of information related to user's opinion widespread in all counties in the world. This paper presents an online system for real-time twitter sentiment analysis and classification. The proposed system helps users to enter the query and get a graphical representation of the tweets polarity. Out of various classification algorithms, Simple Voter and Naïve Bayes algorithms have been used to classify tweets. The obtained results show that the accuracy of the system is efficient using Naïve Bayes classifier.

[5] Title: Using Twitter for Tapping Public Minds, Predict Trends and Generate Value

Authors: Sanchita Kadambari

Description: As the data sets in the world are growing at an exploding rate, research and analysis to derive value from this data has gained ground. Social media is a prime contributor to this data most of which is unstructured. The growing popularity of multimedia and mobile devices has created an overgrowing interconnected network where people are communicating on-the-go, sharing opinions on public forums, blogs, social networking websites. In this context, Twitter has emerged as a major platform to share one's ideas and opinions and millions of users are `tweeting' every second to generate a continuous influx of real-time data. Organizations are tapping into this data to gauge the general opinion or sentiment among the people and methods & techniques are being proposed for sentiment analysis & opinion mining. This has emerged as an active area of research and has a wide range of applications in every field

III. PROBLEM STATEMENT

- Previously for Sentiment analysis, they build the model in opinion mining, for classifying specific words into positive or negative and neutral.
- Opinion mining is also from machine learning concept

3.1. DISADVANTAGES

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

IV. PROPOSED SYSTEM

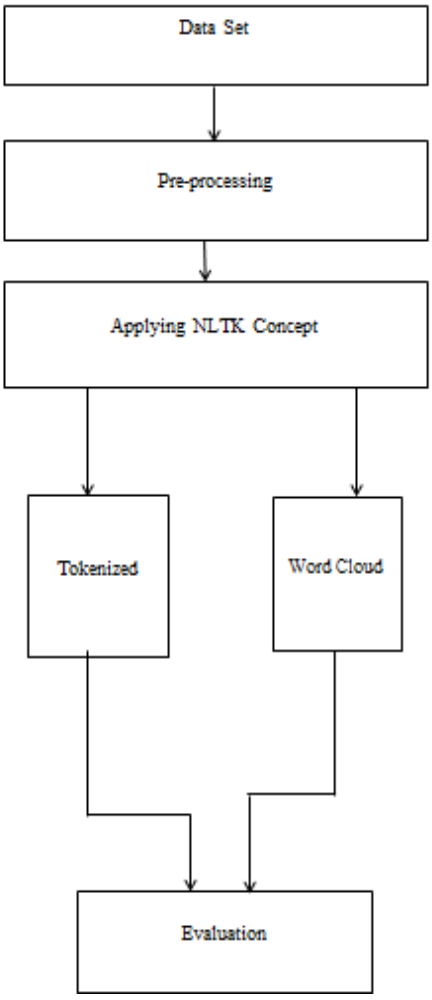
- The proposed model of twitter data analysis will be implemented using Anaconda python.
- The tweets can be analysed and characterized based on the emotions used by the social users. We attempt to classify the polarity of the tweet where it is either positive or negative.

- The data provided comes with emotions, 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 positive and negative which is a form of representation of the "tweet".
- Also, the extra features are added in our project. The Features are to create the world map by using plotly function and also the map are shown in time series prediction in each and every country of the world by using pandas function of the system.

Advantages.

- Better Performance
- High accuracy prediction
- Accurate results

V. SYSTEM ARCHITECTURE



1.5 MODULE DESCRIPTION

- Module 1: Data Collection
- Module 2: Data Pre-Processing
- Module 3: Sentiment analysis
- Module 4: Result

Module 1: Data collection

Extracting real time tweets using Twitter Streaming API For classification and training the classifier we need Twitter data. For this purpose, we make use of API's twitter provides. Twitter provides two API's; Stream API1 and REST API2. The difference between Streaming API and REST APIs are: Streaming API supports long-lived connection and provides data in almost real - time. The REST APIs support short-lived connections and are rate-limited

Module 2: Pre-Processing

In this phase, the tweets are available as text data and each line contains a tweet. Initially we clean up or remove retweets as that will induce a bias in the classification process. We need to remove the punctuations and other symbols that doesn't make any sense as it may result in inefficiencies and may affect the accuracy of the overall process

Module 3: Sentiment analysis

- The sentiment can be found in the comments or tweet to
- provide useful indicators for many different purposes [20].
- Al so, [1 2] and [36] stated that a sentiment can be
- categorized into two groups, which is negative and positive
- words. Sentiment analysis is a natural language processing
- techniques to quantify an expressed opinion or sentiment
- within a selection of tweets [
- The sentiment can be found in the comments or tweet to
- provide useful indicators for many different purposes [20].
- Al so, [1 2] and [36] stated that a sentiment can be
- categorized into two groups, which is negative and positive
- words. Sentiment analysis is a natural language processing
- techniques to quantify an expressed opinion or sentiment
- within a selection of tweets

We build a binary text classifier to classify the sentiment behind the text. We use the various NLP pre-processing techniques to clean the data and utilize the LSTM layers to build the text classifier Implement the NLTK algorithm. The Natural Language toolkit (NLTK) is a library in python, which provides the base for text processing and classification. Operations such as tokenization, tagging, filtering, text manipulation can be performed with the use of NLTK. The NLTK library also embodies various trainable classifiers. NLTK library is used for creating a bag-of words model, which is a type of unigram model for text. In this model, the number of occurrences of each word is counted. The data acquired can be used for training classifier models. The sentiment of the entire tweets is computed by assigning subjectivity score to each word using a sentiment lexicon.

Tokenization

Tokenization is the process by which a large quantity of text is divided into smaller parts called tokens. These tokens are very useful for finding patterns and are considered as a base step for stemming and lemmatization. Tokenization also helps to substitute sensitive data elements with non-sensitive data elements. Natural language processing is used for building applications such as Text classification, intelligent chatbot, sentimental analysis, language translation, etc. It becomes vital to understand the pattern in the text to achieve the above-stated purpose.

word cloud

Word Cloud is a data visualization technique used for representing text data in which the size of each word indicates its frequency or importance. Significant textual data points can be highlighted using a word cloud. Word clouds are widely used for analyzing data from social network websites.

Word Clouds are visual displays of text data – simple text analysis. ʊ Word Clouds display the most prominent or frequent words in a body of text (such as a State of the Union Address). Typically, a Word Cloud will ignore the most common words in the language (“a”, “an”, “the” etc). The remaining words are displayed in a “cloud” with the font size of the word (and-or the colouring of the characters in the word) depicting the relative frequency of occurrence of each target word in the source material. Many freeware packages exist which can process input text and display the resultant Word Clouds.

Module 4: Result

By using sentimental analysis we are classifying the users review by positive, negative and neutral. And we are analyzing and predicting the graph for the project. We have successfully developed python sentiment analysis model. In this machine learning project, we built a sentiment of the tweets into positive and negative.

VI. CONCLUSION AND FUTURE WORK

It is proposed to stream real time live tweets from twitter using Twitter API, and the large volume of data makes the application suitable for Big Data Analytics. A method to predict or deduct the location of a tweet based on the tweet’s information Applying sentimental analysis to extract the sentiment became an important work for many organizations and even individuals. Sentiment analysis is an emerging field in decision making process and is developing fast. Our project goal is to Sentiment Classification system of Twitter Data for Positive and Ukraine sentimental analysis project Negative Reviews of the defined topics. The development of techniques for the document-level sentiment analysis is one of the significant components of this area. Recently, people have started expressing their opinions on the Web that increased the need of analyzing the opinionated online content for various real-world applications. A lot of research is present in literature for detecting sentiment from the text. Still, there is a huge scope of improvement of these existing sentiment analysis models. Existing sentiment analysis models can be improved further with more semantic and common-sense knowledge. A method to predict or deduct the location of a tweet based on the tweet’s information and the user’s information should be found in the future.

REFERENCES

[1] David Zimbra, M. Ghiassi and Sean Lee, “Brand-Related Twitter Sentiment Analysis using Feature Engineering and the Dynamic Architecture for Artificial Neural Networks”, IEEE 1530-1605, 2016.

[2] Varsha Sahayak, Vijaya Shete and Apashabi Pathan, “Sentiment Analysis on Twitter Data”, (IJIRAE) ISSN: 2349-2163, January 2015.

[3] Peiman Barnaghi, John G. Breslin and Parsa Ghaffari, “Opinion Mining and Sentiment Polarity on Twitter and Correlation between Events and Sentiment”, 2016 IEEE Second International Conference on Big Data Computing Service and Applications.

[4] Mondher Bouazizi and Tomoaki Ohtsuki, “Sentiment Analysis: from Binary to Multi-Class Classification”, IEEE ICC 2016 SAC Social Networking, ISBN 978-1-4799-6664-6.

[5] Nehal Mamgain, Ekta Mehta, Ankush Mittal and Gaurav Bhatt, “Sentiment Analysis of Top Colleges in India Using Twitter Data”, (IEEE) ISBN -978-1-5090-0082-1, 2016.

[6] Halima Banu S and S Chitrakala, “Trending Topic Analysis Using Novel Sub Topic Detection Model”, (IEEE) ISBN- 978-1-4673-9745-2, 2016.

[7] Shi Yuan, Junjie Wu, Lihong Wang and Qing Wang, “A Hybrid Method for Multi-class Sentiment Analysis of Micro-blogs”, ISBN-978-1-5090-2842-9, 2016.

[8] Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow and Rebecca Passonneau, “Sentiment Analysis of Twitter Data” Proceedings of the Workshop on Language in Social Media (LSM 2011), 2011.

[9] Neethu M S and Rajasree R, “Sentiment Analysis in Twitter using Machine Learning Techniques”, IEEE – 31661, 4th ICCCNT 2013.

[10] Aliza Sarlan, Chayanit Nadam and Shuib Basri, “Twitter Sentiment Analysis”, 2014 International Conference on Information Technology and Multimedia (ICIMU), Putrajaya, Malaysia November 18 – 20, 2014.