Twitter Data Sentiment Analysis Using ML

C. Hemavathy[1], V Hasitha[2], C Hemasree[3], P Divakar Reddy[4], G Haneef[5]
Assistant professor, Student, Department of CSE, AITS, Tirupati

Abstract

In today's world, social media has become a popular means for people worldwide to share information. Twitter is a platform where users can send and read posts called 'tweets' and interact with various communities. They share their daily lives and express their opinions on various topics such as brands and places. This platform provides a unique opportunity for companies to collect data related to opinions on their products or services. The process of sentiment analysis involves extracting tweets directly from the Twitter API, cleaning and discovering the data, and then feeding it into several models for training. Each tweet is classified based on its sentiment as positive, negative, or neutral

Keywords: Sentiment analysis, social media analytics, Machine learning algorithms, Data mining, Text classification, Twitter sentiment score.

Introduction

Social media platforms such as Twitter,Facebook, and Instagram have revolutionized the way people communicate globally. Users can share their opinions on various products or moments and even influence politics and companies. In this paper, we conducted a sentiment analysis onspecific English tweets related to two popular restaurants, KFC and McDonald's. Our research aimed to determine which of the two restaurantshad a more positive or negative sentiment among Twitter users. We analysed a large number of tweets and classified them as positive, negative, or neutral based on their sentiment. Through this analysis, we aimed to provide insights into the reputation of each restaurant among Twitter users.

Sentimental Analysis has various applications. It is used to generate opinions for people of social media by analyzing their feelingsor thoughts which they provide in form of text. Sentiment Analysis is domain centered, i.e. resultsof one domain cannot be applied to other domain. Sentimental Analysis is used in many real lifescenarios, to get reviews about any product or movies, to get the financial report of any company, for predictions or marketing.

Twitter is a micro blogging platformwhere anyone can read or write short form of message which is called tweets. The amount of dataaccumulated on twitter is very huge. This data is unstructured and written in natural languageTwitter Sentimental Analysis is the process of accessing tweets for a particular topic and predicts the sentiment of these tweets as positive, negativeor neutral with the help of different machine learning algorithm

Related Work

1. Material and Methods:

1. Methods:

This paper focus on mining tweets written in English. We are interested in seeing who people on terms of how good/bad reviews are. Analyzing people’s opinions and what they think about a product from their tweets on social media could bea valuable thing for any business. In our project, we extracted tweets from Twitter using R language. R is a programming language used for statistical computing and machine learning algorithms. In order to extract tweets from Twitter, Twitter API were used to create Twitter application and get authorization. In Rstudio which is an environment and graphical user interface for R, we installed necessary packages and libraries. Some of the packages are (TwitteR, rtweet, R0Auth). By using twitteR package you can extract tweets up to 4000 only [9- 12].

2. Model Building

In this phase, after preparing tweet (removing unnecessary symbols), each tweet was labelled as 1, -1, 0. (That's it: positive, negative, or natural) using unsupervised learning algorithm. Since we do not have pre-classified data, a lexicon-based model used to classify tweets. By using two text files containing a list of positive and negative words, along with more words related to our domain. Each word within each tweet is compared to positive and negative documents in order to find matching words, and classify tweets whether it has more positive or negative words. After that, multiple supervised learning algorithms applied for the purpose of training: Naive Bayes, support vector machine (SVM), maximum entropy, decision tree, random forest and bagging.
* Naïve Bayes: is defined as classifier used to determine the most probable class label for each object.
* Support vector machine: is defined as supervised model, used for classification, regression analysis.
* Maximum entropy: is a classifier used for large variety of text classification.
* Decision tree: are flexible algorithms used to assign label based on the highest score. Random forest: is a supervised algorithm for constructing multiple decision tree.
* Bagging: is a classifier used to taking multiple random samples and use each sample separately to construct a prediction model.

3. Algorithm

Random Forest: A random forest is a machine learning technique that's used to solve regression and classification problems. It utilizes ensemble learning, which is a technique that combines many classifiers to provide solutions to complex problems. A random forest algorithm consists of many decision trees. The (random forest) algorithm establishes the outcome based on the predictions of the decision trees. It predicts by taking the average or mean of the output from various trees. Increasing the number of trees increases the precision of the outcome. A random forest eradicates the limitations of a decision tree algorithm. It reduces the over fitting of datasets and increases precision.

A random forest eradicates the limitations of a decision tree algorithm. It reduces the over fitting of datasets and increases precision.

- It is more accurate than the decision tree algorithm.
- It can generate an acceptable prediction without hyperparameter tuning.
- It eliminates the problem of overfitting in decision trees.Entropy and information gain are the building blocks of decision trees Entropy is a metric for calculating uncertainty. Information gain is a measure of how uncertainty in the target variable is reduced, given a set of independent variables.

4. Result:

In this paper, data extracted directly from Twitter API were used to train and test the models. A lexicon based classifier used a manually created lexicon to find the sentiment of each tweet. Our proposed methodology used a novel approach for using both supervised and unsupervised modeling. As a result, the prediction showed improvements in comparison to existing work where a label data is present. Our model combined several algorithms to get the most fit model for our data. Some metrics were used to validate and test the accuracy of each model [12] as follows.

A. Measurements

- Recall: is defined as number of true positives divided by the number of true positives plus the number of false negatives as indicated in (1).

□□ □(□□+□□)
(1)

- Precision: is defined as the number of true positives divided by the number of true positives plus the number of false positives as indecated in (2).

□□ □(□□+□□)
(2)

- Fscore: is a measure of how accurate a model is by using precision and recall following the formula in (3):

F1_Score = 2 * ((Precision * Recall) / (Precision + Recall))
(3)

A. Cross validation

In cross validation, the original training data set is divided into four groups, 4-fold cross validation for testing and training. After applying validation techniques on the models, the prediction accuracy is found as indicated in Table IV and V .

TABLE IV. ACCURACY RESULT (MCDONALD'S)

McDonald's

Algorithm	Accuracy	Precision	Recall	Fscore	Cross Validate
Naïve Bayes	63%	56%	51%	41%	
SVM	50%	33%	40%	56%	
Maxtent	50%	22%	31%	74%	
Decision Tree	80%	44%	57%	54%	
Random Forest	33%	11%	16%	58%	
Bagging	50%	33%	40%	43%	

TABLE V. ACCURACY RESULT (KFC)

KFC

Algorithm	Accuracy	Precision	Recall	Fscore	Cross Validate
Naïve Bayes	41%	37%	55%	45%	
SVM	67%	67%	67%	60%	
Maxtent	58%	78%	67%	78%	
Decision Tree	55%	67%	60%	54%	
Random Forest	62%	89%	73%	57%	
Bagging	70%	78%	74%	68%	

3. Conclusion:

Sentiment analysis is a field of study for analyzing opinions expressed in text in several social media sites. Our proposed model used several algorithms to enhance the accuracy of classifying tweets as positive, negative and neutral. Our presented methodology combined the use of unsupervised machine learning algorithm where previously labeled data were not exist at first using lexicon- based algorithm. After that data were fed into several supervised model. For testing various metrics used, and it is shown that based on cross validation, maximum entropy has the highest accuracy. As a result, McDonalds is more popular than KFC in terms of both negative and positivereviews. Same methodology can be used in various fields, detecting rumors on Twitter regarding the spread of diseases. For future work, an algorithm that can automatically classify tweets would be an interesting area of research.

4. References:

[1]Proceedings of the 2nd International conference on Electronics, Communication and Aerospace Technology (ICECA 2018) IEEE Conference Record # 42487; IEEE Xplore ISBN:978-1-53860965-1

[2]5th IEEE International Conference on Parallel, Distributed and Grid Computing(PDGC-2018), 20-22 Dec, 2018,
Solan, India

[3]El Rahman, Sahar A.; AlOtaibi, Feddah Alhumaidi; AlShehri, Wejdan Abdullah (2019). [IEEE 2019 International Conference on Computer and Information Sciences (ICCIS).

[4]Chen, Siyuan, Chao Peng, Linsen Cai, and Lanying Guo, "A Deep Neural Network Model for Target-based Sentiment Analysis," International Joint Conference on Neural Networks (IJCNN), pp. 17, IEEE, 2018.

[5]K. Lavanya and C. Deisy. "Twitter sentiment analysis using multi-class SVM," Intelligent Computing and Control (I2C2), International Conference on. IEEE, 2017.

[6]M. Trupthi, , S. Pabboju, and G. Narasimha. "Sentiment analysis on twitter using streaming API," Advance Computing Conference (IACC), IEEE 7th International. IEEE, 2017.

[7]P. Huma, and S. Pandey, "Sentiment analysis on Twitter Dataset using Naive Bayes algorithm," Applied and Theoretical Computing and Communication Technology (iCATccT), 2nd International Conference on. IEEE, 2016.

[8][10] Neethu, M. S., and R. Rajasree. "Sentiment analysis in twitter using machine learning techniques." Computing, Communications and Networking Technologies (ICCCNT), 2013 Fourth International Conference on. IEEE, 2013.

[9]* L. Wasser and C. Farmer, "Sentiment Analysis of Colorado Flood Tweets in R", Earth Lab, 2018. [Online]. Available: https://earthdatascience.org/courses/earth-analytics/get-data-usingapis/sentiment- analysis-of-twitter-data-r/. [Accessed: 01-Mar- 2018].