JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Text Classification Using Different Feature Extraction Techniques

P. Sravani Stanley College of engineering and technology for women, Hyderabad, India
N. Prathyusha Stanley College of engineering and technology for women, Hyderabad, India
Ch. Pavani Stanley College of engineering and technology for women, Hyderabad, India

Abstract - We introduce a novel approach for automatically classifying the sentiment of Twitter messages using NLP and deep learning. These messages are classified as either positive or negative with respect to a query term using sentiment analysis. This is useful for consumers who want to research the sentiment of products before purchase, or companies that want to monitor the public sentiment of their brands. There is no previous research on classifying sentiment of messages on microblogging services like Twitter. We present the results of deep learning algorithm for classifying the sentiment of Twitter messages using distant supervision. Our training data consists of Twitter messages. This type of training data is abundantly available and can be obtained through automated means. We show that deep learning algorithm LSTM which is a variant of recurrent neural network (RNN) have accuracy approximately 80% when trained with this data. This paper also describes the preprocessing steps needed in order to achieve high accuracy. The main contribution of this paper is the idea of using tweets for distant supervised learning.

1. INTRODUCTION

1.1 Introduction

Text classification is the process of assigning tags or categories to text according to its content. It's one of the fundamental task in natural language processing (NLP) with broad applications such as sentiment analysis, topic labeling, spam detection, and intent detection. Unstructured data in the form of text is everywhere: emails, chats, web pages, social media, support tickets, survey responses, and more. Text can be an extremely rich source of information but, extracting insights from it can be hard and time-consuming due to its unstructured nature. Businesses are turning to text classification for structuring text in a fast and cost-efficient way to enhance decision-making and automate processes.

WHAT IS TEXT CLASSIFICATION:

Text classification (a.k.a. text categorization or text tagging) is the task of assigning a set of predefined categories to free-text. Text classifiers can be used to organize, structure, and categorize pretty much anything. For example, new articles can be organized by topics, support tickets can be organized by urgency, chat conversations can be organized by language, brand mentions can be organized by sentiment, and so on. As an example, take a look at the following text below: "The user interface is quite straightforward and easy to use." A classifier can take this text as an input, analyze its content, and then and automatically assign relevant tags, such as UI and Easy to use that represent this text:

HOW DOES TEXT CLASSIFICATION

Text classification can be done in two different ways: manual and automatic classification. In the former, a human annotator interprets the content of text and

categorizes it accordingly.

This method usually can provide quality results but it's time-consuming and expensive.

The latter applies machine learning, natural language processing, and other techniques to automatically classify text in a faster and more cost-effective way.

There are many approaches to automatic text classification, which can be grouped into three

different types of systems:

- Rule-based systems
- Machine Learning based systems
- Hybrid systems

Rule-based Systems: Rule-based approaches classify text into organized groups by using a set of handcrafted linguistic rules. These rules instruct the system to use semantically relevant elements of a text to identify relevant categories based on its content. Each rule consists of an antecedent or pattern and a predicted category. Say that you want to classify news articles into 2 groups, namely, Sports and Politics. First, you'll need to define two lists of words that characterize each group (e.g. words related to sports such as football, basketball, LeBron James, etc., and words related to politics such as Donald Trump, Hillary Clinton, putin, etc.). Next, when you want to classify a new incoming text, you'll need to count the number of sport-related words that appear in the text and do the same for politicsrelated words. If the number of sport-related word appearances is greater than the number of politics-related word count, then the text is classified as sports and vice versa. For example, this rule-based system will classify the headline "When is LeBron James' first game with the Lakers?" as Sports because it counted 1 sport-related term (Lebron James) and it didn't count any politics- related terms. Rule-based systems are human comprehensible and can be improved over time. But this approach has some disadvantages. These systems require deep knowledge of the domain. They are also time-consuming, since generating rules for a complex system can be quite challenging and usually requires a lot of analysis and testing.

1.2 Objective

The objective of this work is to classify the documents based on language, using deep learning algorithm. A wide variety of techniques have been designed for text classification, namely decision tree methods, Rule-based classifiers, LSTM, regression modeling, neural network classifier and so on. In this paper the long short term memory approach and nearest neighbor Classifier are used for Categorization of the Indian language Documents. corpus is created using three south Indian languages such as Kannada, Tamil and Telugu. We will use 100 documents related to cinema of each language and politics.

PROBLEM STATEMENT

In this project I trying to develop a Telegu document classification system using deep learning algorithms with different future selection methods on Telegu news corpus

2. Literature Survey

M Narayana Swamy, M. Hanumanthappa [1] developed "Indian Language Text Representation and Categorization Using Supervised Learning Algorithm" India is the home of different languages. Each state in India has its own official language. The objective of this work is to classify the documents based on language, using supervised learning algorithm. A wide variety of techniques have been designed for text classification, namely decision tree methods, Rule-based classifiers, Bayes classifiers, The nearest neighbor classifier, SVM classifier, regression modeling, neural network classifier and so on. In this paper the decision tree, Naïve Bayes and nearest neighbor Classifier are used for Categorization of the Indian language Documents. corpus is created using three south Indian languages such as Kannada, Tamil and Telugu. We have used 100 documents related to cinema of each language. So, corpus was created using 300 documents. All the documents are cinema related and taken from the WWW.

Madhuri Tummalapalli, Manoj Chinnakotla, Radhika Mamidi [2] developed "Towards better Sentence Classification for Morphologically Rich Languages" In this paper, we present an evaluation of popular deep learning methods for sentence classification on the morphologically rich Indian languages, specifically, Hindi and Telugu. For this purpose, we also created a question classification dataset for Hindi, by translating the TREC-UIUC dataset. We show that character-based input can enhance the performance of current classification systems for morphologically rich languages. Finally, we show that our multiplug-CNN variant is able to perform better than our baselines in two out of three tasks in Hindi and Telugu, while giving comparable results for others. Three of which are for Sentiment Analysis, one each for English, Hindi and

Telugu. The English dataset consists of only two classes (positive and negative), the Hindi and Telugu datasets consist of three classes (positive, negative and neutral) each.

Naga Sudha D, Y Madhavee Latha [4] developed "Comparison of Text Classification Models for Telugu News Articles" In this paper we propose a classification model that supports both the generality and efficiency. It also discusses some of the major issues involved in automatic text classification such as dealing with unstructured text, handling large number of attributes and natural language processing-based techniques, dealing with missing metadata and choice of a suitable machine learning technique for training a text classifier. The experimental results over news articles have been validated using statistical measures of accuracy and Score. The results have proven that the methods significantly improve the performance.

Imran Rasheed, Vivek Gupta Vivek Gupta, Vivek Gupta Haider Banka [7] developed.

3. OVERVIEW OF THESYSTEM

3.1 Existing System

• "Urdu Text Classification: A comparative study using machine learning techniques" This paper consists of an exclusive set of 16,678 Urdu documents of news genre and it carries about 16 different classes as per TREC standards. SVM shows quiet better results than the other classifiers and get 68.73% accuracy, when 80% training data with top 300 attributes are selected whereas on the same parameters decision tree achieved accuracy of only 62.37% The above results were further compared with KNN for different values of k (i.e., 1, 5, and 9)

3.1.1 Disadvantages of Existing System

• The results show very poor accuracy i.e. 51.01%, 54.31% and 55.41% respectively for the same set of parameters. The accuracy of KNN classifier.

3.2 Proposed System

we proposed telugu text document classification using different deep learning algorithms and different feature extraction techniques on telugu corpus. we apply deep learning using lstm approach and, support vector machine, logistic regression using vectorization approach. A corpus is a collection of texts, written usually stored in a computer database. Corpora are the main knowledge base in corpus linguistics. The analysis and processing of various types of corpora are also the subject of much work in computational linguistics, speech recognition and machine translation, where they are often used to create hidden Markov models for part of speech tagging and other purposes. Written texts in corpora might be drawn from books, newspapers, or magazines that have been scanned or downloaded electronically. Feature Selection is the process where you automatically or manually select those features which contribute most to your prediction variable or output in which you are interested in. Having irrelevant features in your data can decrease the accuracy of the models and make your model learn based on irrelevant features

Advantages of Proposed System

Reduces Over fitting: Less redundant data means less opportunity to make decisions based on noise.

• Improves Accuracy: Less misleading data means modelling accuracy improves.

• Reduces Training Time: fewer data points reduce algorithm complexity and algorithms train faster

3.3 Proposed System Design

In this project work, I used five modules and each module has own functions, such as:

1. Dataset
2. Preprocessing
3. Feature Selection
4. Data Prediction

3.3.1 Dataset

Here the dataset for Text Classification that we have used. It contains the data of tweets from The Twitter. Text Classification is a process involved in Sentiment Analysis. It is classification of people's opinion or expressions into different sentiments. Sentiments include Positive, Neutral, and Negative, Review Ratings and Happy, Sad. Here we have given the name to our references.

	Sentiment	ID	Data	Query	User_ID	Text
0	0	1467810369	Mon Apr 06 22:19:45 PDT 2009	NO_QUERY	_TheSpecialOne_	@switchfoot http://twitpic.com/2y1zl - Awww, t...
1	0	1467810672	Mon Apr 06 22:19:49 PDT 2009	NO_QUERY	scotthamilton	is upset that he can't update his Facebook by ...
2	0	1467810917	Mon Apr 06 22:19:53 PDT 2009	NO_QUERY	mattycus	@Kenichan I dived many times for the ball. Man...

3.3.2 *Preprocessing*

Pre-processing is a data mining technique which is used to transform the raw data in a useful and efficient format.

Tokenization: Tokenization is the act of breaking up a sequence of strings into pieces such as word, key words, phrases symbols, and other elements called tokens. Tokens can be individual words phrases, or even whole sentences.

Corpus Cleaning: The data can have many irrelevant and missing parts. To handle this part, data cleaning is done. It involves handling of missing data, noisy data etc.

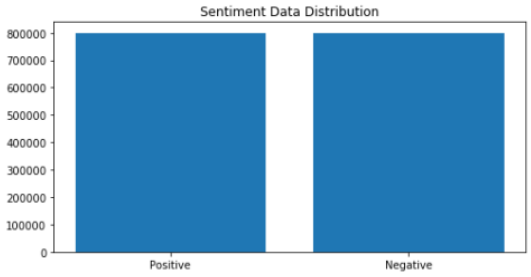
Corpus Transformation: This step is taken in order to transform the data in appropriate forms suitable for mining process.

Corpus Reduction: Since data mining is a technique that is used to handle huge amount of data. While working with huge volume of data, analysis became harder in such cases. In order to get rid of this, we uses data reduction technique. It aims to increase the storage efficiency and reduce data storage and analysis costs.

3.3.3 *Feature Selection:*

Feature Selection is a process of dimensionality reduction by which an initial set of raw data is reduced to more manageable groups for processing. The process of feature Selection is useful when you need to reduce the number of resources needed for processing without losing important or relevant information. We selected features based on two approaches LSTM and Vectorization. The below diagram is for Vector model with feature selection and the algorithm is implemented on a pre-processed data. The pre-processed dataset is split into training set and test set. The training set is used to train the Vector model (Long short-term memory and vectorization) and test set is used to test the model.

3.3.4 *Data Prediction*



	Sentiment	Text
677110	Negative	@B0RNASTARtrell Oh no. IDK if I even wanna know.
1508329	Positive	@BFeld13 I guess just the overall simplicity o...
1219837	Positive	@xscarletmx Pfft, they fail at being as awesom...
1554695	Positive	@riceagain Thanks for that link.. there's a lo...
1166888	Positive	shameless plugging http://rizzysanguinary.devi...
602781	Negative	I don't like being a big ole sneeze face.
699684	Negative	@yamerias Hope U Njoyed the movie - have 2 wai..
539358	Negative	my eyes feel so heavy .
448901	Negative	Just got burnt from my curling iron..

4 Architecture

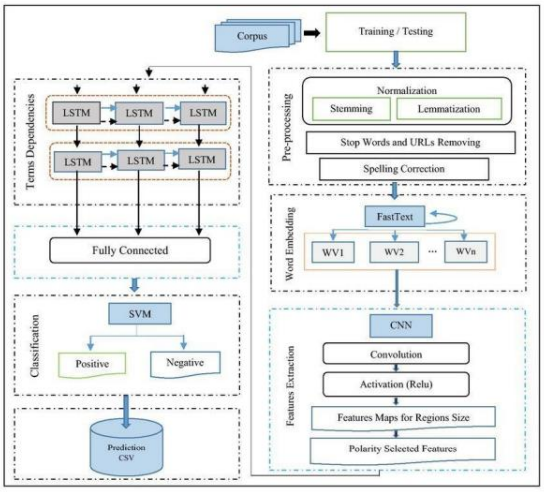


Fig 1: Work Flow Diagram

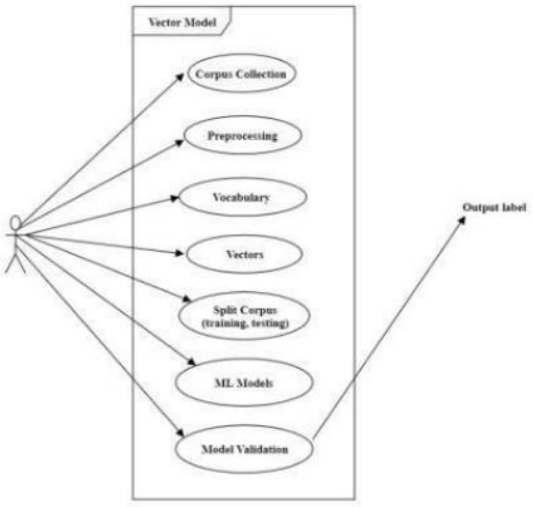


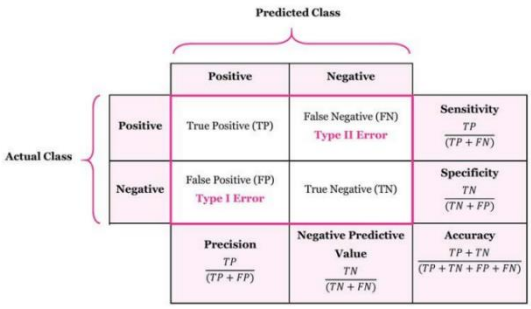
Figure Vector Model

This is where text classification with deep learning steps in. By using text classifiers, companies can structure business information such as email, legal documents, web pages, chat conversations, and social media messages in a fast and cost-effective way. This allows companies to save time when analyzing text data, help inform business decisions, and automate business processes

5 RESULTS ANALYSIS

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.



CLASIFICATION REPORT:

It is one of the performance evaluation metrics of a classification-based machine learning model. It displays your model's precision, recall, F1 score and support. It provides a better understanding of the overall performance of our trained model. Precision: Precision is the ability of a classifier not to label an instance positive that is actually negative. For each class, it is defined as the ratio of true positives to the sum of a true positive and false positive. Precision: - Accuracy of positive predictions. Precision = TP/(TP + FP) Recall : Recall is the ability of a classifier to find all positive instances. For each class it is

6. CONCLUSION

✓ DC Store is a number of information Virtualized file system that targets the consistency of cloud servers in this project. Through client-side data duplication, interior encoding, and a package share management strategy, DC Store not only delivers store high availability, but also cost reductions. When compared to conventional Virtualized storage systems, our small suggested model of DC Store reveals that DC Store dramatically improves quality and cost effectiveness.

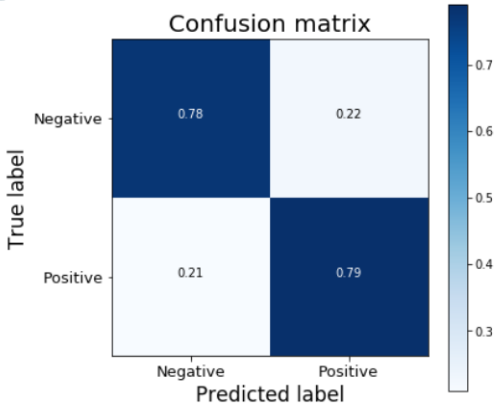
defined as the ratio of true positives to the sum of true positives and false negatives.

Recall: - Fraction of positives that were correctly identified. Recall = TP/(TP+FN) F1

score: The F1 score is a weighted harmonic mean of precision and recall such that the best score is 1.0 and the worst is 0.0. F1 scores are lower than accuracy measures as they embed precision and recall into their computation. As a rule of thumb, the weighted average of F1 should be used to compare classifier models, not global accuracy. F1 Score = 2*(Recall * Precision) / (Recall + Precision)

Support: Support is the number of actual occurrences of the class in the specified dataset. Imbalanced support in the training data may indicate structural weaknesses in the reported scores of the classifier and could indicate the need for stratified sampling or rebalancing. Support doesn't change between models but instead diagnoses the evaluation process.

PERFORMANCE ANALYSIS:



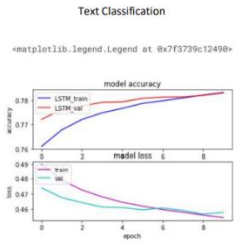


Figure 5.3: Model Evaluation

RESULT

	precision	recall	f1-score	support
Negative	0.79	0.78	0.78	160542
Positive	0.78	0.79	0.78	159458
Accuracy			0.78	320000
Macro avg	0.78	0.78	0.78	320000
Weighted avg	0.78	0.78	0.78	320000

Table : Classification Scores

CONCLUSION AND FUTURE ENHANCEMENT

We will use Deep Learning techniques for implementing and testing models. The system has the capability to classify given text news into three different categories. We are able to achieve satisfactory results based on our training data, which was not available at that moment. In this project we apply Recurrent Neural Network and LSTM(Long Short Term Memory). The Bidirectional wrapper is used with a LSTM layer, this propagates the input forwards and backwards through the LSTM layer and then concatenates the outputs. This helps LSTM to learn long term dependencies. We then fit it to a dense neural network to do classification. I conclude that LSTM is good for some places and make balance the vocabulary both cases. Our Future work includes exploring other classification algorithms on a much more diverse dataset with different Deep Learning techniques. Boosting may also be considered for improving the performance further

7. References

[1] M Narayana Swamy, M. Hanumanthappa "Indian Language Text Representation and Categorization using Supervised Learning Algorithm'', International Journal of Data Mining Techniques and Application, December 2013.

[2] Madhuri Tummalapalli, Manoj Chinnakotla, Radhika Mamidi [2] "Towards better Sentence Classification for Morphologically Rich Languages" Language Technologies Research Center, KCIS, IIIT Hyderabad.

[3] Kapila Rani, Satvika developed "Text Categorization on Multiple Languages Based On Classification Technique" (IJCSIT) International Journal of Computer Science and Information Technologies.

[4] Naga Sudha D, Y Madhavee Latha developed "Comparison of Text Classification Modelsfor Telugu News Articles" Research Scholar JNTUH, Hyderabad may 2018.

[5] K. Pranitha Kumari, A.Venugopal Reddy developed "Syllable n-gram approach for Identification and Classification of genres in Telugu language" 2014 First International Conference on Networks & Soft Computing