International Journal of Computer Applications (0975 – 8887)
Volume 159 – No 1, February 2017

Multilabel Classification of Tweets

Abha Tewari
Mentor
V.E.S.I.T
Mumbai

Pratik Sawant
Student
V.E.S.I.T
Mumbai

Jai Samtani
Student
V.E.S.I.T
Mumbai

Sanket Sawant
Student
V.E.S.I.T
Mumbai

Gaurav Massand
Student
V.E.S.I.T
Mumbai

ABSTRACT
With the help of Social Networking sites many news providers used to share their news headlines on the micro blogging sites such as twitter. We are proposing a system to classify tweets into different groups and labels so that the user can identify the particular tweet from particular category. We will use 120 character tweets for our analysis purpose. Various active and verified twitter accounts would be chosen to extract the tweets. Each tweet is to be classified into 2 category-spam and non-spam. Then further spam group is classified as advertisement, malicious and URL links. The non-spam tweets are classified into 6 labels. These classified tweets then are used to train the various machine learning techniques. Words of each tweet considered as features and a feature vector was created using bag-of-words approach in order to create the instances. The data will be trained using SVM (Support Vector Machine), Naive Bayes and K neighbor machine learning techniques and their efficiency will be compared.

Keywords
SVM -Support Vector Mechanism

NLP -Natural Language Processing

NB-Naïve Bayes

KNN-K Nearest Neighbor

1. INTRODUCTION
Today twitter has grown to be one of the leading social networking sites. From 5000 tweets per day in 2007 to an enormous 500 million tweets per day now twitter user base and subsequent usage has grown rapidly. A large number of these tweets revolve around global happenings around different spheres of life. Twitter has been subject to various analysis and machine learning experiments. In this paper we also propose to perform machine learning on tweets. We are proposing a system to classify tweets into different groups and labels so that the user can identify the particular tweet from particular category. Various active and verified twitter accounts would be chosen to extract the tweets. Each tweet is to be classified into 2 category-spam and non-spam. Then further spam group is classified as advertisement, malicious and URL links. The non-spam tweets are classified into 6 labels. These classified tweets then are used to train the various machine learning techniques. Words of each tweet considered as features and a feature vector was created using bag-of-words approach in order to create the instances. The data will be trained using SVM (Support Vector Machine), Naive Bayes and K neighbor machine learning techniques and their efficiency will be compared. There are several platforms where news is shared. Among all of them Twitter micro blog site was used to collect news as short messages because of following reasons. Many people use Twitter platform to express their views about different topics becoming important source of people opinions. Twitter is everyday growing source as people every day post new news. People from all the age groups and interest use twitter platform. Thus, we can obtain the tweets from twitter of the news that is happening across the globe. The audience from all over the world uses twitter.

The research is to be conducted using the twitter accounts all over the world. Several active and verified Twitter news accounts such as 'BBC', 'Guardian', 'WSJ', and ‘Tech Crunch’ are chosen to extract the data. Twitter API provides the power of fetching n tweets (n should be less than or equal to the number of tweets tweeted by the user) with other necessary details in .csv format. The tweets along with feature vectors will classify the system into 6 groups mainly: Sports, business, health, sports, politics, and entertainment. These groups are chosen because these cover the main areas of a general news for a common person.

2. REQUIREMENTS
The features, elements of the project are required to obtain its requirements. Depending on that requirements are classified into two categories Functional and non-functional requirements.

2.1 Functional Requirements
The functional requirements of the projects are the requirements that should be mandatory included in the system

i.e. the system is incomplete without these requirements. The functional requirements of our system are as follows:

2.1.1Login:
User should be able to login into his twitter account by entering twitter username and password. The credentials will be checked by twitter using TwitterApis.

2.1.2Network Connection
The application should automatically connect to the network to use the network resources

2.1.3 Retrieve Tweets
The application should be able to fetch tweets from twitter related to users account using twitter Library.

2.1.4 GUI
The application user interface should be similar to that of twitter so that user should get similar feel of using application as that of twitter.

2.1.5 Classification of tweet
The user should be able to view the tweets according to particular classes of tweets labels .The classification is done by using machine learning algorithms. The tweets are also classified into groups.

2.2 Nonfunctional Requirements
Non-Functional requirements are the requirements that define the performance, memory requirements, maintenance, and reliability of the system.

2.2.1 Well defined Classification
The classification of the tweets should be as quick and organized as possible.

2.2.2 Label retrieval
The retrieval of the label of classification should be as efficient as possible.

2.2.3 User interface
The user interface should be user friendly and should involve minimum clicks for user to access any resources from the system.

2.2.4 Algorithm Efficiency
We are using various algorithms and comparing them to check which if more efficient for classification of tweets according to labels.

3. CLASSIFICATION
Classification is a function that assigns items in a collection to target categories or classes. The correct prediction of target class for each data item is called classification. Our application aims at Multi-labelled classification of tweets using supervised machine learning algorithms. Tweets are classified into pre-defined classes: Spam Tweets and Non-Spam Tweets.

3.1 Spam Tweets
Spam tweets are the one that violates the rules of twitter policy. Spam is an attempt to mislead or to give an impression to someone ne that the given tweet is authenticated one. Spam tweets are also produced by twitter chat bots.

3.2.1Advertisement Spam:
Electronic spamming is the use of electronic messaging systems to send an unsolicited message (spam), especially advertising, as well ascending messages repeatedly on the same site.

3.2.2 URL Spam:
URL Spam contains malicious urls. By clicking these urls the user will be redirected to another page in which the user may be forced to watch some kind of advertisements or may download some harmful programs which may affect the user's computer.

3.2.3Malicious content Spam
In this type of spam tweet some harmful content will be hidden inside a tweet in the form of url or text. It may contain a shortened URL by clicking it user may be redirected to download some file without even visiting the website.

3.2 Non Spam Tweets
Non-spam tweets are legitimate tweets with proper and correct information. The non-spam tweets are classified into different labels as Health, Technology, Business, Sports, Entertainment and Politics

4. COMPARISON OF ALGORITHMS
The comparison of various algorithms that we will use in the system are as follows:

Table 1. Comparison of Machine learning classification algorithms

Parameter	**NB**	**SVM**	**KNN**
Complexity	Very Simple	Complex	Moderately Complex
Memory	Minimum	Memory Intensive	Memory Intensive
Features	Independent feature Improves performance	Can perform better with dependent features	Can perform good with both
Decision Boundary	Linear/parabolic/elliptic	**Any**	**Any**
Prediction Speed	Fast	Moderate	Slow
Training Speed	Fast	Moderate	Slow

5. TOOLS AND TECHNOLOGIES
5.1 Technologies
The various technologies that will be requiring in the system are as follows:

5.1.1R language:
R language is designed for analyzing the data in the form of graphs which are dynamic and interactive. R is an integrated suite of software facilities for data manipulation, calculation and graphical display

5.1.2 Shiny:
Shiny is an open source R package that provides an elegant and powerful web framework for building web applications using R.

5.2 Tools

The various tools that will be requiring in the system are as follows:

5.2.1R studio

R Studio is free and open source integrated development environment (IDE) for R.

5.2.2 Shiny Dashboard:

For making interactive applications we use shiny dashboard along with R language. The dashboard consists of two files Server. R and Ui.R. Server. R contains all the backend files where Ui.R contains all the GUI part of the files..

5.2.3 Sublime

Sublime Text is across platform source editor with a Python API (API). It natively supports many programming languages and markup languages, and its functionality can be extended by users with plugins.

5.2.4 Tableau

Tableau Software helps to produces a family of interactive data visualization products focused on business intelligence.

5.2.5 Twitter API

Twitter allows you to interact with its data i.e. tweets & several attributes about tweets using Twitter APIs.

5.2.6 UML

The Unified Modeling Language (UML) is a general-purpose, developmental, modeling language in the field of software engineering that is intended to provide a standard way to visualize the design of a system.

6. ARCHITECTURAL DIAGRAM

The Architectural diagram of our system is as follows:

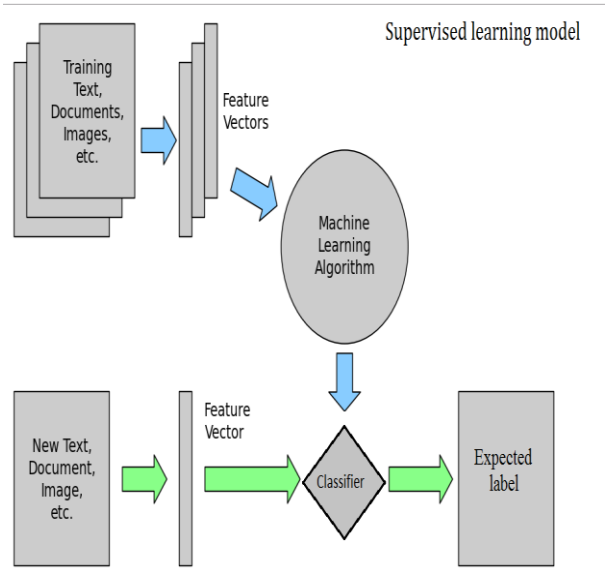

Figure 1. Architectural Diagram

Tweets are obtained in text format using Scripting tools and twitter API. Preprocessing is performed on the tweets. Feature vectors are generated from the words. The algorithm is trained using a part of the dataset. Then it is tested on other part of the data. Bias and variance are evaluated and changes in dataset or feature vectors are made accordingly to fit the data. Once the model is finalized, it is tested against remaining data. Finally it is used for online classification on twitter.

7. STAKEHOLDERS

A stakeholder in the architecture of system is an individual, team, organization, or classes thereof, having an interest in the realization of the system. Right from building the software, testing the software to deploying it and then maintaining it, many people are associated in it. All these people have different requirements and needs from the system. These people are considered as stakeholders in software system architecture. In layman terms, stakeholder is not only limited to technology focused ones (system developer, administrators, testers and support staff), it includes non-technology stakeholders also (such as acquirers, users and communicators). Twitter which is mostly commonly used platform for conveying ideas through text has plenty of users but with limited text capacity. These users basically include athletes, businessman, and president to normal users. This limited text capacity sometimes fails to convey the much valuable information shared through text which makes it necessary to add some type of classification in Twitter. Thus, normal users will have enhanced experience after addition of classification. Many Government employees will also be benefited from this classification as this will make it easier for them to gathering data for some category rather than collecting the whole data and then segregating it manually.

8. DATASET AND IMPLEMETATION

With the advancement of machine learning techniques, now-a-days, many researchers use machine learning techniques for text classification. There are 2 types of machine learning techniques as supervised learning (the learning data in the form of dataset will be provided by the developer) and unsupervised learning (the method will self-learn a clustering procedure by observing the distance among data). For our system we are using supervised learning techniques as our labels do not change regularly.

In order to classify tweets using various machine learning techniques, a proper set of features known as vectors is to be required to extract from the tweets. For extracting tweets bag-of-words approach will be used. The frequency of each word is to be used as data. As there may be large amount of words extracted from feature vectors in different tweets, using all data will cause to increase the overload and dimension. Thus, we will first identify the common words and remove them from the dataset. Once created the dataset, it is important to find a suitable classification method in order to classify the short messages. Support Vector Machine, Naïve Bayes and K nearest neighbor techniques are to be used to classify the data as they capable of dealing with high dimensional dataset. For each subclass we manually found accounts matching the requirements of that particular label for each label we chose over 10verified accounts and fetched 1000 tweets of each account using r studio and twitter library. We saved all the tweets of different classes of each account in different .csv files. Thus in all we gathered 10,000 tweets in all for each sub-class.

These dataset will be used as test cases and will also be used for supervising learning of machine learning algorithms to maximize their efficiency and compare them.

9. ACKNOWLEDGEMENT

We are thankful to our college for considering our project and extending help at all stages needed during our work of collecting information regarding the project. It gives us immense

pleasure to express our deep and sincere gratitude to our project mentor Prof. Mrs. Abha Tewari for her kind help and valuable advice during the development of project and for her guidance and suggestions.

10. CONCLUSION

Thus, we have extracted tweets from different accounts of different classes and will classify them according to their classes. We will train the machine learning algorithms by using dataset as test cases and will compare their efficiency and then the best algorithm as per our outcome will be used for classification of tweets in our system so that user should be able to retrieve a tweet of a particular label in minimum amount of time

11. REFERENCES

[1] ErsinYar,LemiBaruh, Syleyman S. Kozat 2016 Online Text Classification for Real Life Tweet Analysis

[2] P. Selvaperumal, A. Suruliandi 2014 A short message classification algorithm for tweet classification.

[3] InoshikaDilrukshi, Kasun De Zoysa 2014 Twitter news classification: Theoretical and practical comparison of SVM against Naive Bayes algorithm.

[4] Nitin Jindal, Bing Liu 2007 Analyzing and Detecting Review Spam

[5] Shankar Setty ,RajendraJadi , Sabya Shaikh , ChandanMattikalli , Uma Mudenagudi 2014 Classification of Facebook news feeds and sentiment analysis.

[6] KamalanathanKandasamy, PreethiKoroth 2014 An integrated approach to spam classification on Twitter using URL analysis, natural language processing and machine learning techniques.

[7] Support vector mechanism by David Meyer: The interface to libsvm in package e1071

[8] How to Get Started With Machine Learning Algorithms in R by Jason Brownlee: http://machinelearningmastery.com/how-to-get-started-with-machine-learning-algorithms-in-r/

[9] Machine learning course by Andrew Nig: https://www.coursera.org/learn/machine-learning

[10] Basic text mining in r: https://rstudio-pubs-staic.s3.amazonaws.com/31867_8236987cf0a8444e962ccd2aec46d9c3.html