

STUDY OF EFFECTIVE MINING OF UNSTRUCTURED TWITTER DATA

A. Afiya,
Research Scholar
Afiya.aslam.ar@gmail.com

Shaik Javed Parvez
Assistant Professor
parvez.se@velsuniv.ac.in

Dr.S.Arun
Associate Professor
arun.se@velsuniv.ac.in

Abstract:

A large amount of vast data is often termed as Big Data. Most of these data are unstructured and these data are generated from social medium, companies and console records and so on. In a company the potential unstructured data is 80% and above. Mining unstructured big data is the facility of retrieving and extracting useful information from huge datasets. The big quantity and intricacy of unstructured data unlocks much new feasibility for the analyst. Text mining with its method is technique for knowledge discovery from unstructured data. This is a kind of practice applied to unstructured data for data storage and retrieval of required details or information. This paper finds the various data mining approaches and their efficiencies.

Keywords: Big Data, Unstructured Data, Machine Learning, Knowledge Discovery, Twitter, Tweets.

1. Introduction:

Every day data is generated at startling rate. This level of data downpour has excelled our natural ability to understand process, examine, and store these datasets. Let us consider our regular web data. In 1998 Google indexed web pages were around one million but soon it reached around one billion at 2000 and on by 2008 it was one trillion. Recently in 2016, Google indexed web pages were around 130 trillion. This drastic and rapid increase is due to Social Medias like Facebook, Twitter, etc., that allows user to create free contents and store. Web volumes are increased regularly. Further the data generated from mobile phone forms the major breakthrough for retrieving the real data on public from different diverged aspects; mobile carrier can possibly process the large amount of data to advance our daily life. Devices (starting from surveillance camera to cars, to airports) are slackly connected to people. Such devices are connected such that it generates trillion

amounts of mass data. The concept of Knowledge Discovery is the process of retrieving or discovering some valuable information from these dataset which improvises the quality and creates better future living. Data has already reached unexceptional level through Internet of thing (IoT). For example every day before we arrive to office, the system has to process weather, traffic, activity of police to our calendar schedule to optimize the travel time. This level of day to day data generation and optimization led to unstructured big data. In all these actions, we face major challenges in handling large amount of data, including challenges in (a) Capability of system (b) Design of Algorithm and (c) Setting up business model.

This paper focus on big data analytics with comparison and evaluation of different mining techniques and natural language processing for extracting useful information from Big data with most suitable methods. We introduce mining of Unstructured Big Data Chapter 2. Then the summarization of the papers presented related in this issue in Chapter 3, and discuss about proposed system and methodology in Chapter 4 and 5. And we are presenting results at chapter 6 and 7.

2. Big Data Mining : Opportunities and Challenges

Data mining is a process of extracting useful or required information from large dataset. Mostly data comes under three basic categorization such as structured data, semi structured data and unstructured data. An information with high degree of association such that it is easily searchable by any simple straightforward algorithm and addition in relational database is seamless is described as structured data. Semi-structured data lies between the structured and unstructured data. A raw or typed data in a conservative database system is referred as Semi Structured data. These data might not be organized

like table or rows and column, but it is structured data. A lot of data used on the Internet can be illustrated as semi-structured.

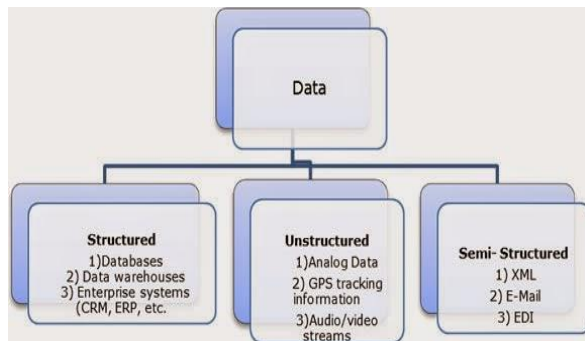


Figure 1 : Categorization of data(Image Source:amitbizintel.blogspot.com)

The nature of Unstructured Data is assorted and it is conflicting. It is a package of various combinations like text, document, image, audio and video. Unstructured data is grows daily at faster rate than structured data. The data which does not fit in traditional database and it is quite opposite to structured data is termed as unstructured data. An example of unstructured data includes email messages, presentation, web pages and other business documents. Though these files have internal structure, it is considered as unstructured because the data doesn't fit orderly in database. It is estimated that 80 percent of organization data is unstructured. And the amount increases significantly than structured data.

IDC forecasts that the global data will grow to 163ZB by 2025. That is ten times the data generated in 2016. Unstructured and Structured data produced by machines, humans are stored in cloud or data center and this data is the new basis for competitive benefit. A data analytic on unstructured data can reveal the inter affiliation among data which was very difficult previously to determine.

3. Related Works:

In this section we analyzed the contribution of four papers towards mining of big data. We also analyze some machine learning algorithm used for mining unstructured data and big data analytics. Data Mining is a process of mining useful information from the database. Various Machine Learning algorithms are used for classification and regressions.

Qingtang Liu et al [1] proposed educational data mining techniques and to integrate inductive content analysis. Sample online post is considered for

inductive content analysis. Based on these results, a single text classification algorithm is applied to classify the sample data. The comparison experiment with SVM showed that this classification has less performance than single naïve bayes classification algorithm on online data. There is a need of more sophisticated algorithm which could take label association into concern to explore online unstructured data.

Ravi Parikh et al [2] employed two unigram models of Naïve Bayes. To classify tweets a Maximum Entropy and a Naïve Bayes bigram model is used. They found that the Maximum Entropy performed comparatively less than Naïve Bayes classifiers. The method of MaxEnt taking advantage of progressive attribute simple doesn't apply well with online post like tweets.

Go et al [3] used the idea of using twitter with emoticons for supervised learning. Training data consist of twitter post with emoticons which serves as noise label. This type of data is available in plenty. They used machine learning algorithms like Naïve Bayes, Maximum Entropy and SVM to build the model and had 80% accuracy with skilled emoticon data. Their characteristic space consists of bigram, unigram and POS and results showed up the performance of SVM outperformed other models. Alexander Pak et al [4] showed how to automatically collect corpus for sentiment analysis. By using collected corpus they build sentiment classifier and classified the tweets on positive, negative or neutral sentiments. In order to collect corpus they retrieved post from famous magazine like New York Times. They used Naïve Bayes Classifier that uses features of N-gram and POS tags

Big data refers to huge volume and complex data that is daily growing dataset from autonomous sources. With rapid development of networking, storage and easy to access the Big data is highly expending in numerous fields like science, engineering and bio medical domains. Xindong Wu et al [5] proposed HACE THEOREM that is Heterogenous, Autonomous, Complex and Evolving theorem. This theorem categorizes the revolution of Big data and propose a model from data mining perspective. This model comprises of demand determined aggression of analysis and mining of source information with security and privacy consideration.

Fatima-Zahra Benjelloun et al [6] presented some Big Data projects in many fields like health care, tourism and politics with examples and models. It conclude traditional technologies are capable of handling Big Data Challenges (complexity, velocity, volume and

variety). Indeed, mining and modeling of Big Data requires more powerful technology. And the another issue is balance between security and privacy. Many developments were made but still it is in downside.

4. Proposed System

There are still many areas which require research for enhancing the potential of Big Data. The proposed system works on framework for data retrieving, storing, analyzing and large scale processing. The major issues faced involving big data are storage and access to the information from huge dataset first challenge faced is storing and accessing the information from the large huge amount of data sets. There is also a big challenge of retrieval of data from social media where new data is generated for every second and constantly changing. Then applying classification algorithms like Random Forest, KNN and Advanced SDA. The main scope of the project is finding the efficiency of each algorithm. To basic method to understand any evaluation includes 3 important aspects like (i) Easy interpretation (ii) Time of Calculation and (iii) Power of Prediction.

KNN is simple yet most commonly used Machine Learning algorithm. It finds application in Data Mining, Pattern Recognition and Intrusion Detection. It is non parametric since it does not rely on underlying assumptions. Random Forest marks the term called ensemble classifier which used in both classification and regression problems. It is collection of trees (decision trees) and it corrects the over fitting of decision tree on training data set. Advanced SDA is based on deep neural network. It takes the input, surpasses it to hidden layers and reconstructs the input at output layer. SDA works like single neural network where it does not depend on predictions of label instead it predicts input at output. The loss is calculated between outputs until loss is minimized.

5. Methodology

5.1 Download Tweets

Consider the most popular social media network called twitter where users share their thoughts through tweets. Twitter API or Tweepy is the platform through which we can mine the data of any user. The user tweets are extracted as data. The first process is retrieval of keys that will help API for authentication. These Keys include consumer secret, access key, consumer key, access level and access key.

Tweepy is cover written in Python for easy access of twitter data. It is one of the library that is installed

using pip. There is a necessity for OAuth Interface to access twitter and to authorize our app. Tweepy is basically open sourced and hosted on GitHub. This application does not rely on user password and it does not work on user password making it more secure. The maximum extraction is limited to 3200 tweets.

5.2 Oppression Tweets Dataset

With the increase in Social Media usages like Facebook, Twitter, Instagram etc education, common issues and solutions, interaction, people voice has increased tremendously there is a thing called cyber bullying is also getting increased. Cyber bullying can cause some serious and bad impact on one person's life. It can pull any one to high level of stress. One solution to stop bullying is automatically detecting the bullying content and block it using machine learning and natural language processing techniques.

Basic requirement to apply machine learning algorithms and test data mining is collecting respective dataset. We need appropriately categorized data where each tweet is categorized as bullying or not. To make apply and test different algorithms for these kind of projects, datasets are created so that students and researchers can apply their algorithms on these dataset and validate results.

5.3 Preprocessing

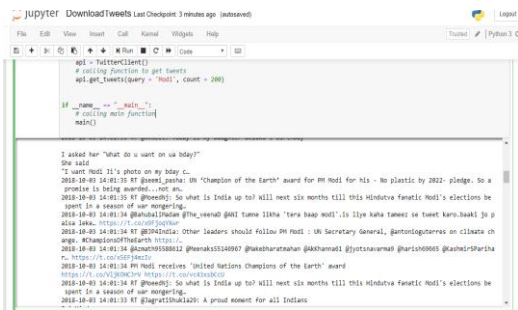
The transformation of raw data into recognizable format is termed as Data Preprocessing in Data Mining. Often real world data is incomplete, inappropriate and may contain errors. Data Preprocessing improves the raw data for further processing so that the dataset includes only required information.

5.4 Training

Pre-processed data needs to be feed into appropriate models for analyzing. Existing methods used Naïve Bayes, SVM and Maximum Entropy algorithm for tweet classification. In our paper we will be categorizing tweets as bullying or not using below algorithms like Random Forest, KNN and SDA. Before data is trained with respective models, the data needs to be split into 70:30 ratio. The 70% data will be used for training and 30% data will be used for testing.

6. Results & Discussions:

This paper provides insight on mining unstructured big data. The Machine Learning algorithm plays major role and contributes in efficient Mining. The process Classification on dataset is important part since it forms the base for big data analytics. Here we are doing data science analysis on various machine learning algorithms and work on the efficiency of each algorithm. In Literature Review we have analyzed that Naïve Bayes form a strong classification model. Through it may exist a lot of advantages there are still some limitations. Further there are other machine learning classification model that needs to analyze for processing.



```

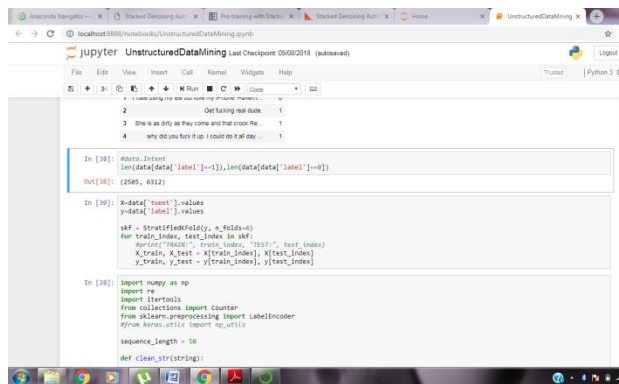
jupyter Download Tweets Last Checkpoint 3 minutes ago (auto)
File Edit View Insert Cell Format Widgets Help
In [1]: from tweepy import API
# calling function to get tweets
api_get_tweets(query = 'Modi', count = 200)

if __name__ == '__main__':
# calling main function
main()

I asked her "what do u want on us today?"
She said
"I want Modi Ji's photo on my body C..
2018-10-09 14:02:16 RT @sneel_pasha: UN "Champion of the earth" award for PM Modi for his - No plastic by 2022- pledge. So a
greatist in India awarded...
2018-10-09 14:02:15 RT @goveerji: So what is India up to? Will meet six months till this Hindutva fanatic Modi's elections be
spelt in a season of war mongering.
2018-10-09 14:02:14 @BabulPrasad @hu_senad @GNI: tune Isha "tera haap mod!" is Isha kaha tawee se tweet karo baaki jo p
sida kaha...
2018-10-09 14:02:14 RT @Dadada: Other leaders should follow PM Modi : UN Secretary General, @anttonoguerres on climate ch
ange. #ModiandOtherLeaders
2018-10-09 14:02:14 @neeraj1348667 @Kaksharanathan @Kakhamati @jyotsnavarshi @harish0666 @kashisharsha
7:
2018-10-09 14:02:14 PM Modi receives 'United Nations Champions of the Earth' award
https://t.co/007056267c
2018-10-09 14:02:14 RT @goveerji: So what is India up to? Will meet six months till this Hindutva fanatic Modi's elections be
spelt in a season of war mongering.
2018-10-09 14:02:13 RT @gopvishvakant: A proud moment for all Indians

```

Figure 2: Downloading Tweets



```

jupyter UnstructuredDataMining Last Checkpoint 09/09/2018 (auto)
File Edit View Insert Cell Format Widgets Help
In [38]: from sklearn.datasets import load_data
load_data(label='1', index=[0:1]), load_data(label='0')
Out[38]: (2585, 6522)

In [39]: X_train, y_train = load_data(label='1')
X_test, y_test = load_data(label='0')

In [40]: X_train, X_test = X_train[X_train.index != X_test.index], X_test[X_test.index]
X_train, X_test = X_train[X_train.index], X_test[X_test.index]

In [41]: import numpy as np
import re
import itertools
from collections import Counter
from sklearn.preprocessing import LabelEncoder
from sklearn.metrics import accuracy_score

sequence_length = 50
def clean_str(string):

```

Figure 3 : Separating Training Data and Test Data

7.Conclusion:

Based on literature survey and results shows unstructured data does not fit on traditional data storage and analysis. But the scope of unstructured data is high and it is necessary to create a single framework for processing the unstructured big data. The volume of data is tremendous issue since it is growing daily. So as volume growth increases the storage and processing become more complex. Most of this industry data goes unused and nature of unstructured data is unverified and trust issues. Another limitation of working with unstructured data

is machine learning conflicts and factor of user behavior. For unstructured data the businesses should come up better way to extract, analyze, and organize the data.

References

[1] Qingtang Liu, Si Zhang, Qiyun Wang, and Wenli Chen. "Mining Online Discussion Data for Understanding Teachers Reflective Thinking" on IEEE transaction 2017

[2] Ravi. Parikh and M. Movassate, "Sentiment Analysis of User-Generated Twitter Updates using Various Classification Techniques", Final Report, CS224N.,2009

[3]A. Go, R. Bhayani, L.Huang. "Twitter Sentiment Classification Using Distant Supervision". Technical Paper , Stanford University 2009

[4] Alexander Pak and P. Paroubek. "Twitter as a Corpus for Sentiment Analysis and Opinion Mining". In Proceedings of the Seventh Conference on International Language Resources and Evaluation, 2010,

[5] Xindong Wu, Xingquan Zhu, Gong-Qing Wu, Wei Ding, "Data Mining with Big Data", IEEE transactions on Knowledge and Data Engineering, Vol.26, No.1 , January 2014.

[6] Benjelloun, Fathima-Z. , Ibn Tofail University., Kenitra, Morocco, Lahcen, A.A., Belfkih, 5., "An Overview of Big Data Opportunities, Applications and Tools", IEEE Conference on Intelligent Systems and Computer Vision, March 2015.

[7] Matthew Herland, Taghi M Khoshgoftaar and Randall Wald, " A Review of data mining using Big Data in Health Informatics", Springer Journal of Big Data, 2014.

[8] Wei fan and Albert Bifet "Mining big data: current status, and forecast to the future" ACM SIGKDD Explorations Newsletter, 2013

[9] Sowmya R, Suneetha K R. "Data Mining with Big Data", 2017 11th International Conference on Intelligent Systems and Control (ISCO),2017

[10]Algorithm understanding from www.analyticvidhya.com

[11] Big Data reference from www.insidebigdata.com