Record Linkage And Data Prediction Using Big data

[1]Prof.Vishal Shinde, [2]Miss.Dhanashri Kamatkar, [3]Miss.Pratiksha Partole, [4]Miss.Neha Patil, [5]Miss.Aishwarya Waghchoude.

[1]Asst.Professor,[1,2,3,4,5]UG Student,[1,2,3,4,5]Computer Engg. Dept. ShivajiraoS.Jondhle College of Engineering & Technology, Asangaon, Maharashtra, India.

[1]mailme.vishalshinde@gmail.com,[2]dhanashrikamatkar1997@gmail.com,[3]pratikshapartole777@gmail.com,[4]patil.neha6178@gmail.com,[5]aishwaryawaghchoude96@gmail.com

Abstract-The process of text organization in which the text documents which having words, phrases and word permutation with respect to set of previously defined categories. Text organization having many applications which can be the classification of content, filtering email news contracted-casting and monitoring. Extracted keywords from documents for classification of the documents. Keywords can be a subset of words which having the most important information about the content in document[1]. The important keywords from document has to be extracted using this method. The research has to be done using K-Nearest Neighbor (KNN) and Naive Bayes algorithms its act are analyzed.

Keywords: Data security,Privacy Preserved RecordLinkage,Data Mining,WordNet, IDF,KNN.

I. INTRODUCTION

Most of these connections are made by textual messages, which makes a hard task for physical impaired people to speak with others. Thus, one method which treats this problem that got interest by investigators is called word prediction. It's a word processing characteristic that aims decline the digit of keystrokes essential for typing words[1]. One of the input method in text mining is Text classification to categorized the documents in an supervise manner. This method can be used for routing and document filtering to topic specific processing mechanism such as machine translation and information removal. Various Methods are used for classification using Support Vector machine, Naive bayes, K-Nearest neighbor and Decision trees. In the proposed work, using TF-IDF and Word Net the keywords are extracted from documents. From each document there are incomplete number of words are selected. Documents are classified using machine learning method According to the extracted keywords, specific processing mechanism such as machine translation and information extraction. Various methods are used for classification of document such as naive Bayes algorithm.Using TF-IDF the keyword extracted from documents.Record Linkage is data connectivity .In the circumstance of record linkage categorization refers to the process of dividing record pair into match and non-match.

II. LITERATURE SURVEY

Data mining is used for mining a data into the databases and also finding out meaningful patterns from the database Implementation data mining has been carry out for a variety of application over a classification algorithms which is having large collection and for varying data size. There contain many possible variants; some of them can be discussed in following section.

Paper 1: Menaka S* Research Scholar PSGR Krishnammal Prison for Women Peelamedu, Coimbatore: Wongkot Sriurai[1]compared the quality processing method of Bag of Words (BOW) with the topic model. Text categorization algorithms such as support vector technology (SVT), naive bayes (NB), and decision tree are used for testing.

Paper 2: M.Priya, Dr. R. Kalpana and T. Srisupriya Spell checking defines the several methods for error correction and spell checking. It also delivers the information about many existing spell checking systems which developed for several Indian languages. They described two method for spell checking is (1) Statistical technique used for N Gram Analysis (2) Dictionary lookups. Baljeet[3] the part of error detection and spell correction method has been surveyed. With the support of spell checkers the Existed task associated in Punjabi and the Punjabi language is also considered. With more accuracy, the Punjabi spellchecker will be implemented by using edit-distance and dictionary lookup based method.

Paper 3: Shahidul Islam Khan;Abu Sayed MD Knowledge finding from various data sources requires the merging of attention data.

© 2019, IJREAM All Rights Reserved.

III. COMPARTIVE STUDY

Table No.3.1 Comparative table

Sr No.	Paper Title	Author's Name	Technology	Advantage	Disadvantage
1.	Health Data Integration with Secured Record Linkage.	Shahidul Islam Khan; Abu Sayed Md. LatifulHoque Dept of Computer Science & Engg(CSE) Bangladesh University of Engg and Tech (BUET) Dhaka, Bangladesh	A novel based detection system for permission analysis in android system is proposed.	Used to protect mobile phone user privacy	Decompilation based permission lead unexpected problem..
2.	Text classification using Keyword Extraction method.	Menaka S* Research Scholar PSGRKrishnammal Prison for Women Peelamedu,Coimbatore	Hybrid approach for android security.	Need for secure ways to detect,control and avoid malicious behavior has become high	. Statement analyse and dynamic analyse helps to prove false positive from static analysis not be true.
3.	Hybrid Model For Word Prediction Using Naive Bayes.	Henrique X. Goulart, Mauro D. L. Tosi, Daniel Soares-Gonc¸alves, Rodrigo	J48 classification Algorithm was used to detecting android leaking sensitive data.	Detect malwares and analyzing source code or permission	The main problem when information in mobile application try to access by android permission
4.	Naive Bayes model with enhanced Negation managing and n-gram method for outlook classification	Sumit Kumar Garg Ronak Kumar Meher	Navy	It's relatively simple to understand and build It's simply skilled, even with a little data set It's fast!	It assumes every attribute is autonomous which isn't always the box

IV. PROBLEM STATEMENT

Works with organized and not well organized records, and thus process of extracting when the sources are not well organized or semi-organized. Elobrate concept simulations to deal with missing, conflicting, and corrupted dataset.The major advantage of the method described above is that, this method cannot be used for any primary key. Utilizes non-matching ,record linking information in count to direct matching.

V. PROPOSED SYSTEM

Records in data bases are arrogated to represent find of articles occupied from a individual population. The records are assumed to contain some attributes (fields or variables) verifying an individual entity. Examples of identifying attributes are id, name, address ,mobile no, age and gender.

Positives True (PT) - These are the properly expected progressive values which means that the value of real and the output of expected is also yes.

Negatives True (NT) - These are the properly expected negative values which means that the output of real class is no and value of predicted class is also no.

Positives False (PF) – When real class is no and expected Class is yes. Example, if real class says this student did not seat but expected class tells they that this student will seat.

Negatives False (NF) – When real class is yes but expected class in no. E.g. if real class value indicates that this student seated and predicted class tells they that student will weak. Once they recognize these four parameter then it will be find Precision, Recall, Accuracy and F1 score.

- **Accuracy**

Accuracy = PT+NT/PT+PF+NF+NT

- **Precision**

Precision = PT/PT+PF

- **Recall**

Recall = PT/PT+NF

- **F1 score**

F1 Score = $2* \frac{(Recall * Precision)}{(Recall+Precision)}$

VI. ALGORITHM

Algorithm 1: Training Algorithm

```
Require: M,N
M: Set of Documents (Training dataset)
N: class of the document positive ,negative
1: V= Extract feature Vector (A)
2: D = Number of training documents
3: for c in C do
4: Dc= Number of documents with class c
5: prior[c] = Dc/ D
6: for w in V do
```

© 2019, IJREAM All Rights Reserved.

```
        likelihood[w][c]=(count(w,c)+k)/((k+1)*Dumber of words in class c)
    end for
end for
return prior, likelihood
```

Algorithm 2: **Testing Algorithm**

```
Require: t
t: Document to test
V = Extract Feature Vector(t)
for k in K do
    score[k] = prior[k]
    for w in V do
        score[k]= score[k] * likelihood[w][c]
    end for
end for
return argmax(score[c])
```

VII. MATHEMATICAL MODEL

(i)**K-Nearest Neighbor***:* The k-nearest neighbor algorithm is used to test the degree of law of similarity between documents and k training data.

$$R^x \le R_{kNN} \le R^x \left(2- \frac{MR^X}{M-1}\right)$$

(ii)**Naive Bayes Algorithm*:*** Naive Bayes classifier is a easy probabilistic classifier based on applying Bayes Theorem with independence assumptions. The more expressive term for the underlying quantity model would be independent.

$$M(x=u|D_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}} e^{-\frac{(u-m_k)^2}{2\sigma_k^2}}$$

(iii)**N-Gram analysis:** In this, N-gram Analysis Method is used. N-grams are n-character sub arrangement of words or strings where n is set to one, two or three. One character n-grams are termed as uni-grams, two character n-grams are termed as bi-grams, and three character n-grams are termed as tri-grams.

VIII. SYSTEM ARCHITECTURE

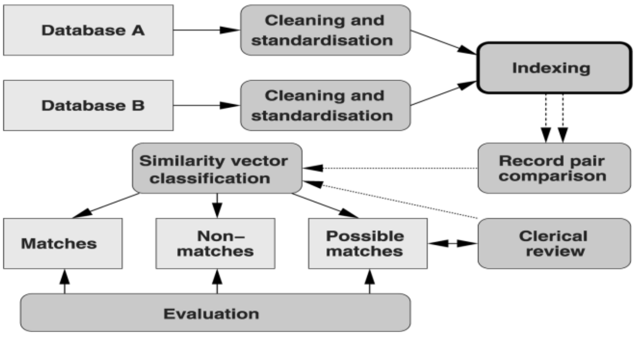

Fig.8.1: System Architecture

It consists of three separate subsystems. First is the data giving out subsystem that takes the submitted text, extract n-grams and calculates possibilities from the number of occurrence of the n-gram. In second part in which the classifier subsystem are uses the base possibilities to analyse the possibilities of the occurrence of each word in the circumstance of the goal words. The last Subsystem's query is frequency approximation and it uses the trouble-free Decent-Turing method to analyse the possibility of previously invisible words. The particular most important article of the structure is that all these three Subsystems are totally coupled. The classifier can be totally altered and only the format of storing the last possibilities must remain the same. That makeing the subsystems can be system.

IX. ADVANTAGE

1)It is Fast.
2)It is Very Simple to Understand.
3)It can be easily train on small dataset.
4)It is highly scalable algoritham.

X. CONCLUSION

Thus we have tried to implement paper on [2] Menka S and Radha N, IJARCSSE 2013 International Journal of Advanced Research in software Engineering.Text Classification using keyword Extraction Method .The input insight is that new method for modelling text from Information Retrieval, for performing similarity joins from Database Research and for learning optimal decision boundaries from data mining can potentially improve the record linkage process in terms of manual effort and measurability. The record linkage in machine learning models out perform the possibilities in Record Linkage model with admiration to the exactness and the completeness metrics, the possibilities in record linkage model identifies a minor percentage of possibly corresponding record pairs, both the Clustering and the Mixture Record Linkage models are very useful mostly in the case of real applications where training sets are not available or are very lavish to gain

REFERENCE

[1] Bo Pang and Lillian Lee. A Sentimental Education, Using subjectivity summarization based on sentiment analysis to minimum cuts. In Proceedings of the 42nd annual meeting on Association for Computational Linguistics, page 271. Association for Computational Linguistics, 2004.

[2] Janyce M Wiebe, Rebecca F Bruce, and Thomas P O'Hara. Development and use of a gold-standard data set for subjectivity classications.In Proceedings of the 37th annual meeting of the Association for Computational Linguistics on computational linguistics, pages246{253}

[3] Minqing Hu and Bing Liu.Mining and summarizing customer reviews. In Proceedings of the Tenth ACM SIGKDD international conference on Knowledge discovery and data mining, pages sd168{177}ACM, 2004.

[4] Fuchun Peng and Dale Schurmans. Combining naive Bayes and n-gram language models for text classication. Springer, 2003.

[5] Vivek Narayanan, Ishan Arora, & Arjun Bhatia. speedy and accurate emotion classication using an improved Naive Bayes model. In Intelligent Data Engg. & Automated Learning{IDEAL 2013, pages 194{201} Springer, 2013.

[6] Sanjiv Ranjan Das and Mike Y Chen. Yahoo! for amazon Sentiment parsing from small talk on the web. In EFA 2001 Barcelona Meetings, 2001.

International Journal for Research in Engineering Application & Management
IJREAM