MJRET
Open Access

Multidisciplinary Journal of Research in Engineering and Technology

Journal homepage: www.mjret.in

ISSN:2348 - 6953

PIPELINED FRAMEWORK FOR ANALYZING IDENTITY THEFT BEHAVIORS USING TEXT MINING

Neha Sharma, Prof. Sunil Damodar Rathod
Computer Department.
Dr. D. Y. Patil School of Engineering,(Affiliated to Savitribai Phule Pune University)
Pune, India

Abstract: *In today's era, identity theft is the most growing issue in society, businesses where the aspects of lives are digital. Identity theft is gateway crime, as criminals use others identities to steal the money for an Example credit card number, hack the social network authentication, etc. identity theft behaviors are identified by the various government investigative agency or media sources. These investigation committees are analyzing the identity theft records or news of identity thieves, which is in the form of recorded stories, report by media, rich site summary news feeds and hypertext markup language over the internet. At the National Institute of Justice "Identity theft becomes the defining crime of the information age, estimated amount of incident per year is 9 million or more. Over the past decade, the Federal Government and most states have passed legislation to impose criminal sanctions on identifying thieves.*

In this paper we present the pipe line system to better analysis of identity theft behavior by mine the identity theft stories using text mining technique .In this research news stories and records are collect from internet by using keywords related to identity theft, these records or stories are filter using text mining techniques and eliminates the duplicate information from the records and generate well formatted record in simple text format .By using this system can analyze pattern and resources used by thieves to commit identity theft.

Keywords*: Identity theft, Identity theft records, Text mining, PII attributes.*

1. INTRODUCTION

Earlier, the "identity theft "was little used and known term. But today's it is a widely used term, one associated with an event that has caught by public, media and government attention and which is becoming a serious social issue.[1]

The narrow definition of identity theft, typically enroll by the industry participants, describes this behavior as the assumption of a person's identity and the use that identity and associated personal data to make new credit accounts. [3]

The main purpose design pipeline framework is provided to analyze records of identity theft to the better understanding pattern of identity thieves. In these system basic three components as follows:

1) ***News provider***: Collect news records or stories of identity theft from internet news API.
2) ***Filter manager***: Filter the news stories by applying mining by keywords such as bank related news, stock market related news.
3) ***News Validation Manager***: In this phase duplicate copy of news records is eliminated.

2. RELATED WORK

Individual's identity is a collection of personal detail that can include, name, address, driver's license, social security number, telephone number, place of employment, employee ID, mother's maiden name, account numbers, or CC number (Federal Trade Commission, 2005). There is a possible way to define how identity theft occurs. First, identity theft can occur when a victim opens an account or uses the personal information of a victim. In the second step, the identity thief deflects the billing statements of the accounts. In This step the victim is not aware about stealing identity from him or her. The third step is when the overdue payments result in a poor credit rating. The purpose of this research is to design is to develop a repository of relevant knowledge to better understand the processes how identity thieves and fraudsters. Commit identity thefts. To understand the criminals' process, that allows the crime to take place, provide resources, and how we can be preventing it. We hope this work will decrease identity theft and fraud vulnerabilities. The information extraction from raw text format has been approached by many research efforts. However, Use of this technique in the identity theft research area is relatively more. The existing system is based on the Identity Threat Assessment and Prediction (ITAP) project at the Center for Identity at The University of Texas at Austin .Related works of existing system are the natural language processing research and text mining studies. Text mining is the process of gleaning useful information from natural text format. The goal is to analyze the text from news stories and extract meaningful data for the understanding of identity theft behaviors.

The natural language text format is unstructured and it's difficult to understand by systems so text mining usually used for transforming the natural language text into a well-structured format, detecting lexical and syntax phrase, and evaluating and analyzing the data. General text mining research includes text categorization, entity extraction, sentiment analysis, entity relation modeling, and so on. Text mining techniques help process large amounts of unstructured data, such as biomedical applications, social network applications, marketing applications and sentiment analysis.

The Next is, text mining in stock market price prediction. The main purpose of this study is to investigate how text mining techniques can be used to predicting the future trends of stock prices by analyzing news articles (breaking news). There articles published that is related stock market prediction like stock price, volumes, company income, company cost, and so on, instead of textual information from news articles. It is important to investigate how stock markets react to breaking news because if we know this we can create fast computerized systems that automatically analysis new news articles before the market has had time to adjust itself to the new information. This research was making much more profit on stock trades. [3]

3. PROBLEM IN EXISTING SYSTEM

To predict identity threats, ITAP tool is developed by the University of Texas need amount of data on which they can perform the analysis and generate better result but there is a problem that is there is no public repository provide updated identity theft and fraud. Thus, a lack of identity theft data is in proper structure is a problem Solution to these problems is some government organizations have internal databases record identity theft and fraud, but these are publicly available. Another solution is sourced from Internet (e.g. News articles posted by media), which post identity theft news stories every day. These articles are in a simple raw text form. We can’t analyze it directly. To solve this problem by proposing an automated solution that uses text mining, an application of natural language processes, to extract the meaningful information from the identity theft stories and articles. This information helps to extract patterns of identity theft behaviors and to prevent identity theft crimes in future. This research provides source which is publicly available and that helps to inform businesses, consumers, agencies, and researchers about the processes of identity theft.

4. PROPOSED SYSTEM ARCHITECTURAL FLOW

This proposed system is framework for analysis of identity theft behaviors using text mining. This system provides the news records in well-structured text format with meaningful information about identity theft.

In this system first collect identity theft related news records by using keywords related to identity theft from the internet then filter the news records by using text mining technique such as tokenization. In this technique preprocessing of new stories to eliminate the irrelevant information by using name entity recognizer. Named Entity such as name of person, location, time, etc. In next, preprocessed records are used for validation process which checks the duplicate copy of news records. If records already exist in our local database, then it discards those particular records otherwise generate identity theft records with meaningful information. We hope this record is helpful for government agency or businesses.

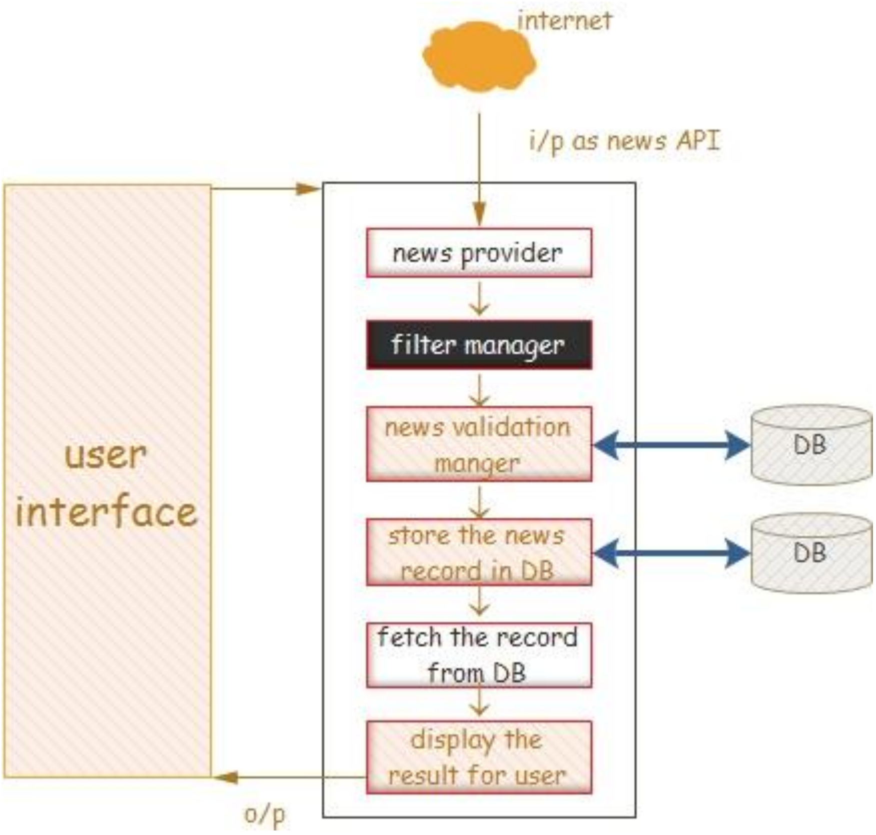


Fig.1: Process flow Architecture

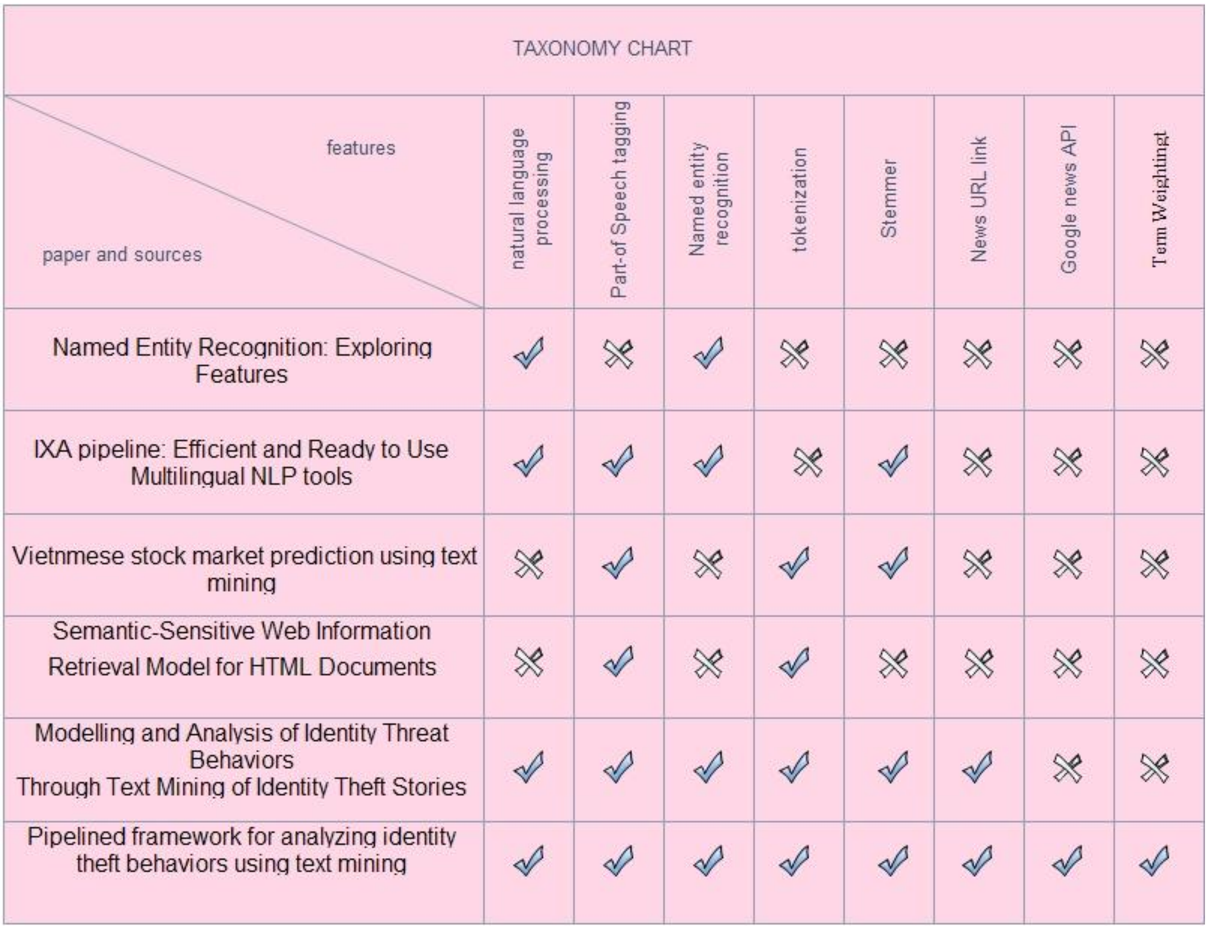
Let us introduces text mining techniques commonly used by text mining systems; these are also planned for the proposed system.

1) ***Text Preprocessing**:* preprocessing, use to eliminate the language dependent factors, so that the structure becomes clearer. One of the most common preprocessing Techniques is Tokenization used for text preprocessing. It defines processing of splitting a text stream, example a sentence, into tokens, phrases, words, symbols or other elements. Stemming process is used for inflected words into a base form so that the number of phrases or similar meaning words can be reduced. For example, word ‘look’ can be altered with a structural suffix to produce similar words such as ‘looks,’ ‘looking,’ and ‘looked.’ These words all share the stem ‘look.’ It is usually beneficial to map all inflected forms into the stem. Most commonly used stemmer is porter stemmer.It also complicated for such exception case word like ‘be’, ’was’ and ’see’ etc.

2) ***Named Entity Recognition**:* In this technique name entity is referred as phrases that include the Names of persons, locations, expressions of the times, monetary values, and so on. For example: Alice comes from japan at 8:00 pm. This sentence contains three named entities: “Alice” is a person, “japan” is a location and “8:00 pm” is time. Named entity recognition (NER) is an important process for information extraction.

3) ***Part-Of-Speech Tagging**:* The process of assigning a part-of-speech to each word in a sentence. Example each word is what a noun, verb, phrase etc.

5. TAXONOMY CHART



TAXONOMY CHART

features paper and sources	natural language processing	Part-of Speech tagging	Named entity recognition	tokenization	Stemmer	News URL link	Google news API	Term Weightingt
Named Entity Recognition: Exploring Features	✓	✗	✓	✗	✗	✗	✗	✗
IXA pipeline: Efficient and Ready to Use Multilingual NLP tools	✓	✓	✓	✗	✓	✗	✗	✗
Vietnmese stock market prediction using text mining	✗	✓	✗	✓	✓	✗	✗	✗
Semantic-Sensitive Web Information Retrieval Model for HTML Documents	✗	✓	✗	✓	✗	✗	✗	✗
Modelling and Analysis of Identity Threat Behaviors Through Text Mining of Identity Theft Stories	✓	✓	✓	✓	✓	✓	✗	✗
Pipelined framework for analyzing identity theft behaviors using text mining	✓	✓	✓	✓	✓	✓	✓	✓

Fig.2: Taxonomony chart

6. CONCLUSION & FUTURE SCOPE

In this paper presents a of concept for collecting and analyzing publicly available News stories assess and predict the behaviors of identity thieves and fraudsters. Our research work used text mining techniques on news articles collected from the Internet to analyze and understand criminal behaviors and predict future possibilities of identity theft and fraud. The first step in the system involves the collection of News Stories from the Internet. We take news records from internet from these few Records May be invalid for further process. Subsequently, the text of these stories Were Preprocessed, eliminating irrelevant and unnecessary information. Then, named entities were extracted using the named entity recognizer. This record can be used to conduct the analysis about different aspects of the identity theft, such as the risk of exposure for a particular PII attributes, the identity theft in different market sectors, the location based identity theft. This analysis aims to increase actual knowledge about identity threat behaviors, offer early warning signs of identity theft, and thwart future identity theft crimes.

ACKNOWLEGEMENT

We thank our guide Mr. Sunil Damodar Rathod for his inspiration and guidance.

REFERENCES

[1] Graeme R. Newman and Megan M. McNally, "Identity Theft Literature Review," National Institute of Justice, 2005.

[2] AWARE Software, Inc. "Awareness User Guide," AWARE Software, Inc. Austin, Texas, 2014.

[3] Aase, Kim-Georg, "Text Mining of News Articles for Stock Price Predictions," Norwegian University of Science and Technology, 2011.

[4] Julia S. Cheney "Identity Theft: Where Do We Go From Here" A Discussion Forum Sponsored by the Payment Cards Center and the Gartner Fellows Program.

[5] Wang, Yi, and Xiaoping Wang, "A new approach to feature selection in text classification," Machine Learning and Cybernetics 6, 2005: 3814 -3819.

[6] Boiler pipe, "boiler pipe: Boilerplate Removal and Full text Extraction from HTML pages." [Online]. https://code.google.com/p/boilerpipe/,March, 2014.

[7] https://en.wikipedia.org/wiki/Identity_theft.