Finding Substituted Word using Viterbi Algorithm

Sonal S. Deshmukh
Assistant Professor
Department of MCA
MGM's Jawaharlal Nehru Engineering College,
Aurangabad, Maharashtra, India

Abstract— **Finding original meaning of any substituted word is really a challenge in text mining. Mostly criminals or terrorist are using different techniques to achieve their goal. Word substitution technique can be used by them to work secretly or to hide actual text from others in their documents. Various algorithms using different measures can be used to detect this substitution. In this paper I implemented Viterbi algorithm to improve the detection rate. 20 News Group dataset is used for implementation and testing. 87.8% of detection rate is achieved by this method.**

Index Terms— **Hidden Markov Model, Viterbi, Stemmer, Stopwords**

Introduction

As the technology is growing, cases of cheating, fraud cases are also increasing. Internet is most common technology used by people not only for their official but also for personal work. Easiest communication means like the internet chatting systems and the emails are the best and the most common for the maintaining contacts with the people around the world. Unfortunately many people make misuse of this technology. They use these applications for illicit purpose like criminal activities, fraud activities, bribe etc. Internet can be use by many terrorist organizations and their supporters for a wide range of purposes including recruitment, financing, propaganda, training and gathering and dissemination of information for terrorist purposes. Use of the Internet for terrorist activities creates both challenges and opportunities in the fight against terrorism. Terrorist wanted to work secretly and try to hide meaning of information in their documents. So they replace sensitive or harmful words by normal words so that other person cannot recognize it. For e.g. Indian Muzahideen was using word 'Musa' instead of word 'Gujarat'. In a sentence "Collect some people for work from Gujarat", if Gujarat is replaced by Musa, then sentence would be "Collect some people for work from Musa". Likewise terrorist may try to use some code to hide the information. Apart from the messages, the terrorist groups are using sites to publish objectionable material like method to prepare Bomb etc [1].We tried to detect such substitution by using Hidden Markov Model technique.

I. LITERATURE SURVEY

There are various ways to hide the information from others. Normally people feel that encryption is the only method used for hiding any communications from surveillance. They thus tend to concentrate on just decryption techniques applied to the intercepted communications [2]. However, there are millions of emails moving around the Internet every day. In such a large numbers, it is very easy for subversive groups to send their illicit communications without encryption and the chances of detection are very rare. In fact, in such cases, the senders are well aware of the fact that encryption may actually attract attention to a mail which might otherwise have passed unnoticed. Hence, such mails are hardly ever encrypted by the criminals. Hacking is one of the most important misuse of internet It includes sending text messages via email or SMS to the group members either using fake identification or by hacking/stealing the device or network link. Another way is Phishing. This is a process where the culprit poses as an official entity and tries to acquire important financial or personal details like credit card number, social security number, home address or telephone number.

Internet extortion is a serious crime done by the criminals. Threatening emails exploiting people and blackmailing them on the internet to pay money or other favors is another identity theft, child pornography, password trafficking, spamming, virtual stalking, fake auctions and other internet scams are some of the threatening criminal activities that exist on the Internet.

Another method to hide the data is deceptive writing. Deception refers to the act of deceiving, misinforming or misleading. [3] Shows a method for detecting stylistic deception in written documents. This method is used to detect misuse of author's name to any document. Also deceptive writing is used to make fraud by changing the original data to hide the meaning of information.

In order to hide the meaning of the messages written, the illicit groups started substituting the word in the message by a normal word and the message looks like a normal mail. Earlier terrorist group used to adopt some common techniques for substitution, example of which can be substitution by replacing the word with the word having near wordcount rank e.g. the word "missile" have word rank 6316 and "garment" have word rank 6319 [4][5].
Different measures were used by researchers to detect this substitution. But we tried to improve the detection rate by using Hidden Markov Model in which possibility of finding hidden state can be improved.

II. VITERBI ALGORITHM

A Markov Model is a system which produces Markov chain and hidden Markov model is a system where the rules of producing the chain are hidden or unknown. First rule include the probability that there will be a certain observation and second rule include the probability that there will be a certain state transition, given the state of the model at certain time. It is a mathematical approach for solving many types of problems. Hidden Markov model is a statistical tool for modeling a wide range of time series data. . Mathematical theory of Markov process is written by Andrei Markov hence name given Markov Model in early twentieth century [7]. To solve various problems mainly three algorithms are developed.

- Forward-Backward
- Viterbi
- Baum-welch

In this research work, Viterbi algorithm is considered since it is used to find most likely hidden sequence of the states called as Viterbi path. A practical application was developed where author implement a forward error correction decoder using Viterbi algorithm which shows its relation to continuous observable variables in probabilistic system [8].

Viterbi algorithm is given as below.

Suppose we are given a Hidden Markov Model (HMM) with state space S, initial probabilities π_i of being in state i and transition probabilities $a_{i,j}$ of transitioning from state i to state j. Say we observe outputs $y_1, y_2, \ldots, y_T$. The most likely state sequence $x_1, \ldots, x_T$ that produces the observations is given by the recurrence relations.

$$V_{1,k} = P(y_1|k).\pi_k$$
$$V_{t,k} = P(y_t|k).max_{x \in S}\left(a_{x,k}.V_{t-1,x}\right) \qquad (1)$$

Here $V_{t,k}$ is the probability of the most probable state sequence responsible for the first't' observations that has 'k' as its final state.

The Viterbi path can be retrieved by saving back pointers that remember which state 'x' were used in the second equation. Let $Ptr(k,t)$ be the function that returns the value of 'x' used to compute $V_{t,k}$ if $t > 1$, or k if $t = 1$. Then:

$$x_T = argmax_{x \in S}(V_{T,x})$$
$$x_{t-1} = Ptr(x_t, t) \qquad (2)$$

Here we are using the standard definition of arg max. The complexity of this algorithm is $O(T * |S|^2)$.

III. PROPOSED METHODOLOGY

This work uses 20 News Group Dataset containing around 80000 News. Following assumptions and setup is used for application of the algorithm.

1. Hidden States: There are two states which need to be decided. They are whether the statement contains substitution or not. Accordingly two states NR (not replaced) and R (replaced) were considered i.e. {NR, R}.
2. Probabilities of hidden states: Whether the statement contains substitution or not is equally probable. Hence the probability of occurrence of each state at start is considered same i.e. {0.5, 0.5}.
3. Observables: In the experimentation the observables are the statements under consideration and eventually the constituents of the statement i.e. words in the statement.
   In order to get the observables and its probabilities, all the News in 20 News Group dataset are processed as below
   a. Every News is processed for stemming using porter stemmer.
   b. Stopwords in the News were kept as it is as their occurrence may have an impact in calculation of the probabilities.
   c. The frequency data obtained is used to generate the frequency matrix for entire News Group.

Thus observables and its probabilities are obtained.

4. Transition Probabilities: In order to obtain the state transition probabilities, a list of sensitive word was prepared and its substitutions were also decided. Then another instance of 20 NewsGroup Corpus was created by making the copy of original News Group corpus folder. In the second folder, the sensitive words were replaced in some files by its substitution. If News is selected for substitution, all the sensitive words were substituted in that News to ensure that substitution is to hide the meaning of the entire News data.

   Here substitution is done in 45% of files as that can be the maximum amount of substitution and the most critical to detect. If the substitutions are less in amount, the percentage difference of probabilities of original and substitution is more and hence it is easy to detect. However, 45% substitution has given the behavior around similar like the original one. Hence we selected this number to test the results of Viterbi for worst condition.

   This selection helped us in defining NR→NR transition probability as 0.55 and NR→R transition probability as 0.45, considering that there are only two states. Similarly, transition probability of R→R transition is 0.45 as that many percent of files are substituted and R→NR transition probability as 0.55.

5. However, as it is to be decided whether the statement is substituted or not, there is no need to consider the transitions from one state to other as it is needed to find maximum of the probabilities for NR transition and R transitions i.e. maximum of probability p of producing sequence only form NR and probability p of producing sequence only form NR.

   For some observable sequence ($O_1O_2…O_n$), find the probability p for following sequence NR, NR,…., NR and R,R,….,R. Other sequences containing NR and R transitions need not be considered.

6. Once the p(NR,NR…,NR) and p(R,R,….,R) is calculated, then the state of the given statement is:

$$S = \begin{cases} NR & if\ p(NR, NR, \dots, NR) > p(R, R, \dots, R) \\ R & Otherwise \end{cases} \quad \textbf{(3)}$$

IV. Performance Analysis

Adopting above mentioned steps, a state diagram for Viterbi algorithm for text substitution detection problem is designed which can be seen in the figure 1 given above.

The probabilities are calculated depending on the word frequencies. Table I shows the sample word count for subset of files before replace and after replace.

Probability of occurrence of a word, p(w) in the corpus is calculated as,

$$p(w) = \frac{\sum_{n=1}^{n=N} d_w(n)}{\sum_{m=1}^{m=M} \sum_{n=1}^{n=N} d_{w_m}(n)} \quad \textbf{(4)}$$

Where $d_w(n)$ is a frequency of the word ‘w’ for document ‘n’. N represents the number of documents available and M represents the number of word in the corpus.

Probabilities thus calculated are used in calculation Viterbi algorithm for further calculation.

For testing purpose, various words are selected for substitution and accordingly statements were classified as NR and R. Also some statements are constructed using the words in the corpus which have a meaning and some substitution is decided for it. Based on the probabilities listed in Table II below and using viterbi variant designed above, a probability of whether the statement belongs to substituted class i.e. R class or NR class is calculated. Whichever probability is more, a decision is taken for labeling of that statement in R or NR.

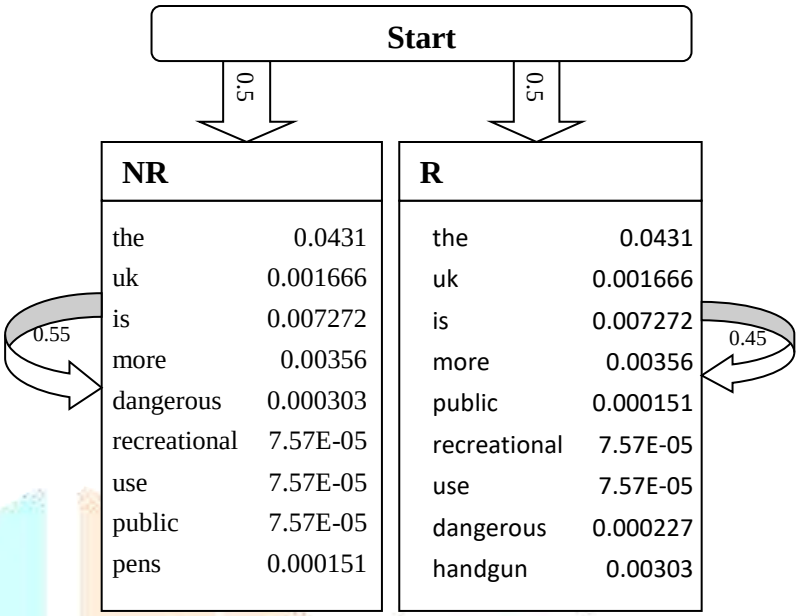

Fig 1 : HMM Model for Text Substitution Detection in 20News Group Corpus

TABLE I:
Word frequencies from sample table before (after) substitution, used for calculation of probability for NR state.

Words	File A	File B	File C	File D	File E	File F	File G	File H
Handgun	4(0)	15	2(0)	25			6(0)	
Pen	1(5)		2				6	
Death		8					3(0)	
Birth							1(4)	
Murder		3	1(0)				1	
Call			1					
Weapon		13	1		5(0)		1	
Paper					2(7)			
Violence		4(0)	1					
News		1(5)						
Attack		11			2(0)			
Friendship					2			
Gun		23	9	43	1	1	2	
Football								

Consider an example statement “the UK is more dangerous”. It is to be decided here whether this statement is substituted or not. In order to decide this let us first find out probability that the statement belongs to not substituted or NR class.

It is already discussed that being both the hidden states are equally probable, π value will be 0.5. Other occurrence probabilities are as given in the table. Probability of having the next word from same class i.e. NR and R are 0.55 and 0.45 respectively.

Considering all the values, probability that the statement “the UK is more dangerous” is not substituted i.e. it belongs to NR class is:

p_{NR}("the UK is more dangerous")= π_{NR}*p_{nr}("the")*$p_{nr\to nr}$ * p_{nr}("UK")*$p_{nr\to nr}$ p_{nr}("is")* $p_{nr\to nr}$ * p_{nr}("more")*$p_{nr\to nr}$ * p_{nr}("dangerous")
Putting the values,
p_{NR}("the UK is more dangerous")= 0.5*0.04309953*0.55 * 0.00166641*0.55*0.00727163* 0.55* 0.00356007*0.55* 0.00030298
= $2.57743*10^{-14}$
Similarly, for calculation of probability that statement belongs to R class,
p_R("the UK is more dangerous") = π_R*p_r("the")*$p_{r\to r}$ * p_r("UK")*$p_{r\to r}$ p_r("is")* $p_{r\to r}$ * p_r("more")*$p_{r\to r}$ * p_r("dangerous")
= 0.5*0.04309953*0.45 * 0.00166641*0.45*0.00727163* 0.45* 0.00356007*0.45* 0.000227238
= $8.66257*10^{-15}$
This is as per rule given in (3).
It can be concluded that the given statement is Not Substituted and belongs to class NR.

TABLE II: Probabilities for words

Word	Total Count	Probability
The	569	0.04309953
Uk	22	0.00166641
is	96	0.00727163
more	47	0.00356007
dangerous	4	0.00030298
recreational	1	7.5746E-05
Use	1	7.5746E-05
Of	59	0.00446902
handguns	52	0.0039388
minimally	2	0.00015149
restricted	2	0.00015149
public	1	7.5746E-05
Pens	2	0.00015149

The method is used for various statements. Some samples are as given below in the Table III. This method is used to test around 700 different statements. And it was found that this method has classified **87.8%** instances correctly.

TABLE III: Probabilities Of Sample Statements

Statement	Probability of Belonging to NR	Probability of belonging to R	Class
the uk is more dangerous	2.57743E-14	8.66E-15	NR
the uk is more public	6.44356E-15	5.78E-15	R
the use of handguns is minimally restricted	1.32729E-22	3.06E-23	NR
the use of pen is minimally restricted	5.10497E-24	9.19E-24	R
the uk use handguns	1.78253E-12	7.51E-13	NR
the uk use pens	6.85589E-14	2.25E-13	R

V. CONCLUSION

This work focuses on detection of substitutions in communication which is used to hide the meaning of the sentence. HMM algorithm can be used to detect the substitutions. By using two NR and R classes, we can find the probability of substituted sentence. Virterbi algorithm find more suitable for detection and using this model has resulted in accuracy of 87.8% for 20 News Group dataset.

REFERENCES

[1] Hsinchun Chen, Edna Reid, Joshua Sinai, Andrew Silke, Boaz Ganor, "*Terrorism Informatics: Knowledge Management and Data Mining for Homeland Security*", Springer 2008 , pp 431

[2] Mrs. Shilpa Mehta, Dr. U Eranna, Dr. K. Soundararajan, *"Surveillance Issues for Security over Computer Communications and Legal Implications"*, Proceedings of the World Congress on Engineering 2010, WCE 2010, June 30 July 2, 2010, London, U.K

[3] Afroz, S. ; Dept. of Comput. Sci., Drexel Univ., Philadelphia, PA, USA ; Brenan, M. ; Greenstadt, R. *"Detecting Hoaxes, Frauds, and Deception in Writing Style Online, Security and Privacy(SP)"* Symposium on 20-23 2013 San Francisco, CA, ISSN :1081-6011 pp 461-475.

[4] www.wordcount.org/main.php

[5] Szewang Fong, Dmitri Roussinov, And David B. Skillicorn, "*Detecting Word Substitutions In Text*", IEEE Transactions On Knowledge And Data Engineering, Vol. 20, No. 8, August 2008

[6] Lawrence R. Rabiner, "A *Tutorial on Hidden Markov Models and selected Applications in Speech Recognition"*, proceeding on IEEE (vol 77, issue 2) Feb 1989.

[7] A. Markov. *"An example of statistical investigation in the text of Eugene onyegin, illustrating coupling of tests in chains."* Proceedings of the Academy of Sciences of St. Petersburg, 1913.

[8] G. William Slade, *"The Viterbi Algorithm Demystified"*, ResearchGate Publication, 3/2013.