Content-based Analysis in order to Detect Sensitive Data

Nazım DÖLEKÇEKİÇ – Alper PUNAR
Batuhan TATLISERT – Beril YOKARIBAŞ
Dr. Nurdan SARAN
Çankaya University, Department of Computer Engineering

Abstract

Companies or organizations must store and protect all personal information about their customers, employees, and all other written communications against any inner or outer security threats.

There are numerous things that can threaten security. For instance, personal information sent by e-mail, information or image in a certain document, screenshots of bank account so on and so forth. We can analyze data with Data Leakage Prevention (DLP) systems and increase security against any data leakage. DLP systems are used to preserve sensitive or personal data, and protect against unauthorized disclosure.

In this project, we survey the techniques that can be used to statistically analyze data in motion within DLP systems and how to intercept leakage of sensitive data.

Introduction

The main purpose of this project is to detect documents, and texts, with sensitive information using content-based analysis. The system includes an E-mail Relay and model.

The scope of the project is to use this system to detect sensitive information in e-mails. It will check the mail's text and its PDF attachments. All text in the PDF including images will be converted to plain text and the model will decide whether it contains sensitive information or not. We used machine learning to construct the model.

Solution

Data that contains sensitive information will be given to the system so that it can understand what makes that data sensitive. Data can be from any topic and area but for better results, it would be better to constrain it so the success of determining whether it is sensitive or not would be higher.

We used the data from NSA and various agencies as secret data and used the data from NSA and NATO news as the public data so the public and secret data contents would be more correlated. We constructed the model with these data sets.

Later we put the model to the filter part of the E-mail Relay we implemented.

An e-mail will come to the E-mail Relay and it will enter the filter part which includes our model.

It will give a score to it, which indicates whether it has secret content or not. If it has a score lesser than 0.5 it will be forwarded to its target destination. If it has a higher score than the threshold it will be stopped and its contents will be sent to the admin for further examination. Admin will decide to send or stop it indefinitely.

We used Labse for transfer learning using TensorFlow. To extract the images from PDFs we used Apache Tika and Tesseract.

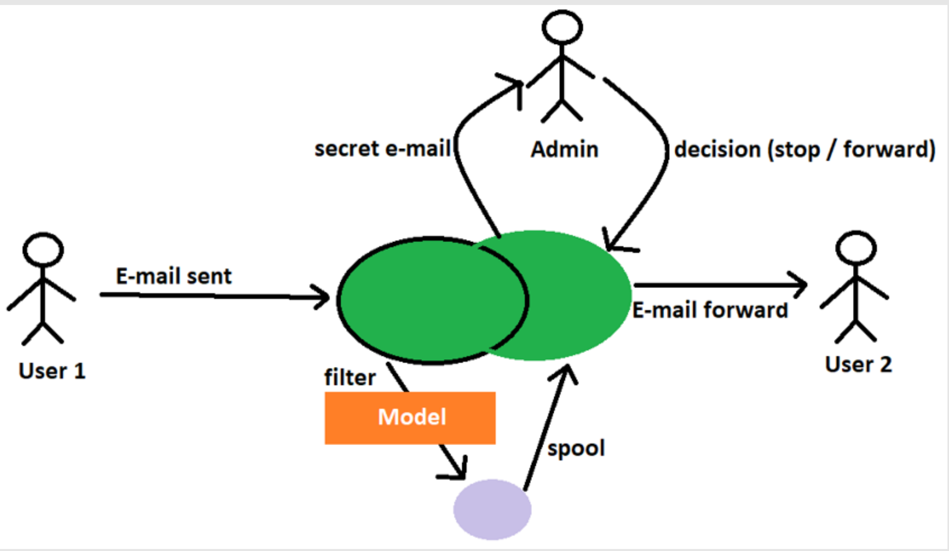
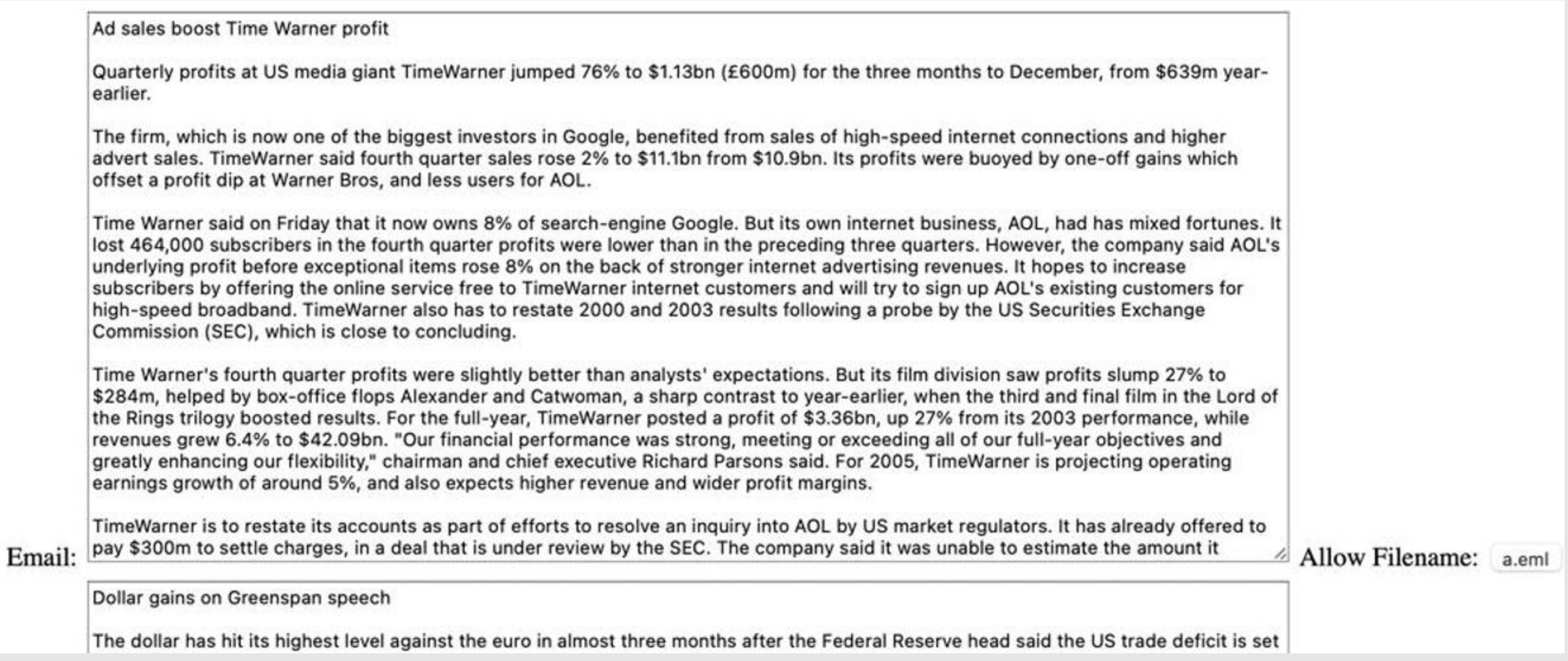


Figure 1 – General Flow



Email:

Ad sales boost Time Warner profit

Quarterly profits at US media giant TimeWarner jumped 76% to $1.13bn (£600m) for the three months to December, from $639m year-earlier.

The firm, which is now one of the biggest investors in Google, benefited from sales of high-speed internet connections and higher advert sales. TimeWarner said fourth quarter sales rose 2% to $11.1bn from $10.9bn. Its profits were buoyed by one-off gains which offset a profit dip at Warner Bros, and less users for AOL.

Time Warner said on Friday that it now owns 8% of search-engine Google. But its own internet business, AOL, had has mixed fortunes. It lost 464,000 subscribers in the fourth quarter profits were lower than in the preceding three quarters. However, the company said AOL's underlying profit before exceptional items rose 8% on the back of stronger internet advertising revenues. It hopes to increase subscribers by offering the online service free to TimeWarner internet customers and will try to sign up AOL's existing customers for high-speed broadband. TimeWarner also has to restate 2000 and 2003 results following a probe by the US Securities Exchange Commission (SEC), which is close to concluding.

Time Warner's fourth quarter profits were slightly better than analysts' expectations. But its film division saw profits slump 27% to $284m, helped by box-office flops Alexander and Catwoman, a sharp contrast to year-earlier, when the third and final film in the Lord of the Rings trilogy boosted results. For the full-year, TimeWarner posted a profit of $3.36bn, up 27% from its 2003 performance, while revenues grew 6.4% to $42.09bn. "Our financial performance was strong, meeting or exceeding all of our full-year objectives and greatly enhancing our flexibility," chairman and chief executive Richard Parsons said. For 2005, TimeWarner is projecting operating earnings growth of around 5%, and also expects higher revenue and wider profit margins.

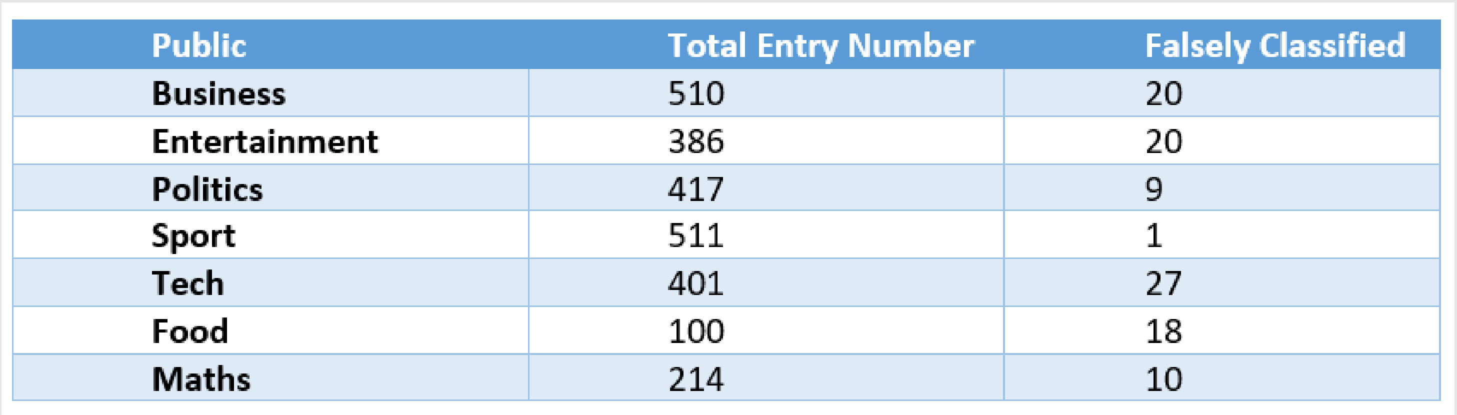
TimeWarner is to restate its accounts as part of efforts to resolve an inquiry into AOL by US market regulators. It has already offered to pay $300m to settle charges, in a deal that is under review by the SEC. The company said it was unable to estimate the amount it

Allow Filename: a.eml

Dollar gains on Greenspan speech

The dollar has hit its highest level against the euro in almost three months after the Federal Reserve head said the US trade deficit is set

Figure 2 – Admin Mail Interface



Public	Total Entry Number	Falsely Classified
Business	510	20
Entertainment	386	20
Politics	417	9
Sport	511	1
Tech	401	27
Food	100	18
Maths	214	10

Figure 3 – Results

Results & Conclusion

Thanks to both 407 and 408 we learned the importance of literature review, Software Requirements Specification, and Software Design Documents and familiarize ourselves with project development stages. We also gained knowledge regarding data security and Data Prevention Systems as well as machine learning.

For secret data testing, we used the separated secret data from the main data set, 620 entries, to test the model. There were 6 falsely classified entries in that test.

Acknowledgement

We would like to thank our valuable project supervisor Nurdan SARAN. She answered our questions tirelessly throughout the Project and supported us.