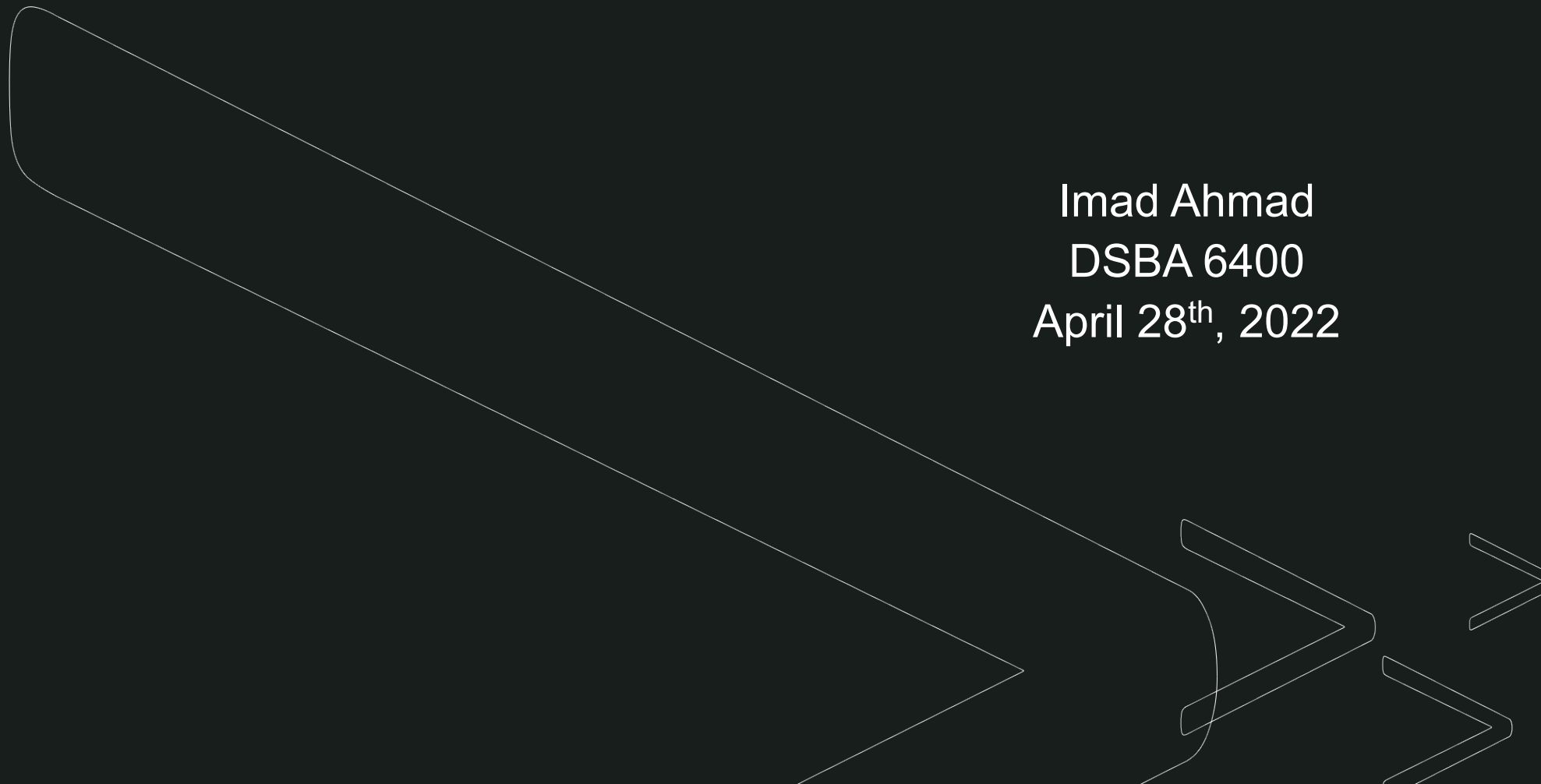
Internal Audit: Using Natural Language Processing to Capture IT Material Weaknesses

Imad Ahmad

DSBA 6400

April 28th, 2022



Leverage Natural Language Processing for Information Extraction

- **Project Goal**:

  1) Build a tool which utilizes NLP for information extraction.

  2) Extract relevant sentences which contain IT Material Weaknesses from Internal Audit reports.

  3) Classify IT Material Weaknesses into categories so that companies can create a strategy to mitigate risks posed.

- **Data**:

  271 Internal Audit reports spanning the years 2015 – 2020

Gain a Competitive Advantage by Using Natural Language Processing

SAVE TIME & RESOURCES:
Manual extraction of information is time consuming and cumbersome.

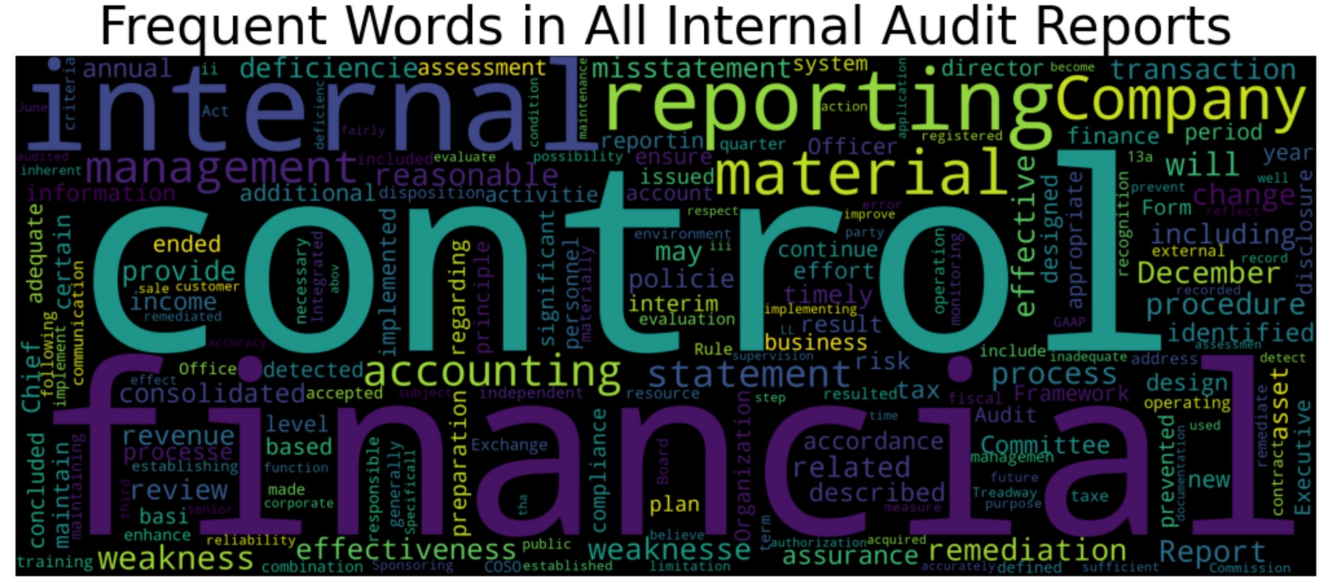
EXTRACT COMPLETE AND ACCURATE INFORMATION:
NLP libraries can help find specific information through phrase matching and word similarity.

Mitigates the risk of missing valuable information.

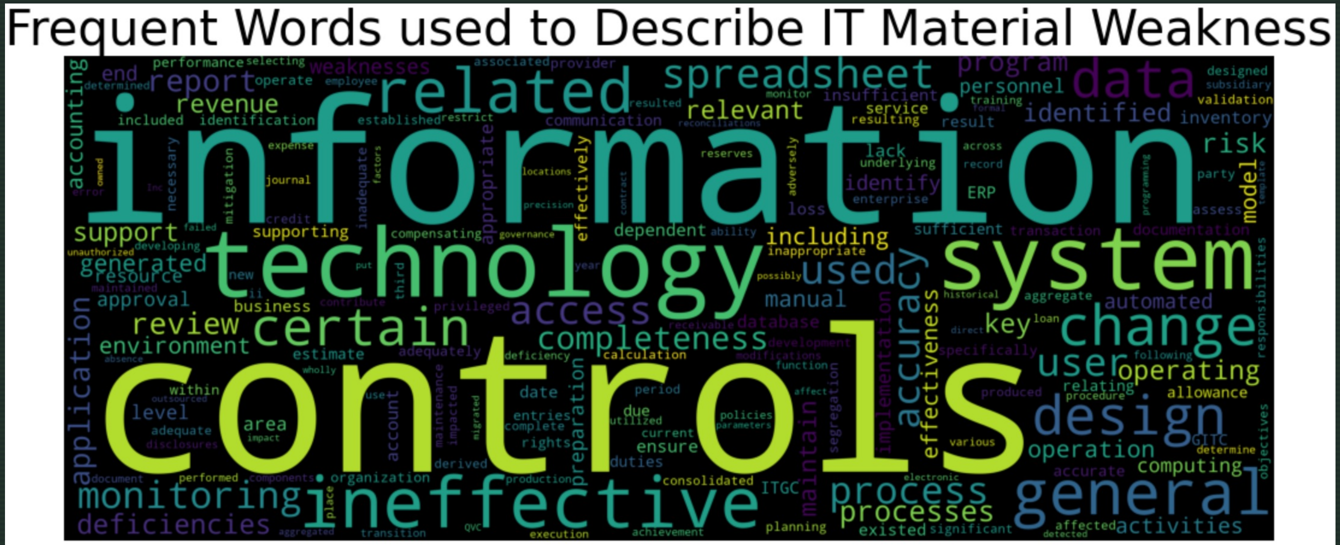
DISCOVER NEW / HIDDEN INFORMATION:
NLP can help find new information OR discover information that is hidden within the text.

Glance into the Data:

- Frequent terms used across all audit reports.



- Frequent terms used to describe **IT Material Weakness**



Tool Performance: Discovering New Information

- **Compare:**
  Ground Truth vs. Extracted information
  * Ground Truth: collected via manual extraction

The Tool Extracted Relevant Information	27 Cases of New Information
100 % of the time.	

- **Manually Validate New Information:**
  Cases where the tool extracted information when there was no ground truth

Tool Extracted Information is Detailed but Concise

ground_truth	extracted_text
lack of Information Technology governance	physical safeguards and access to programs/data We have inadequate design of IT general and application controls that could prevent the information systems from providing complete and accurate information consistent with financial reporting objectives and current needs. network infrastructure systems have been safeguarded and a help desk system has been implemented to monitor significant events in 2015.
Ineffective general information technology controls over user access, program change and system development which resulted in ineffective segregation of duties and automated controls throughout the organization.	We did not maintain effective general information technology controls (GITCs) which restricted users' access to application system. databases and operating systems throughout the organization that are used for financial reporting purposes. establishing general information technology controls to assign appropriate user access to the information technology operating system. systems of control may not prevent or detect all misstatements. even effective systems of control can provide only reasonable assurance of achieving their control objectives.

- **NEW INFORMATION**

ground_truth	extracted_text
none given	[certain persons at such subsidiaries had access to conduct conflicting accounting operation. , we are in the process of reviewing our employees' access throughout our accounting systems., We also plan to deploy new information technology tools in order to improve our control over the segregation of accounting duties.]

Classification

- Group similar IT material weaknesses into categories
- Help companies create strategies to mitigate risk

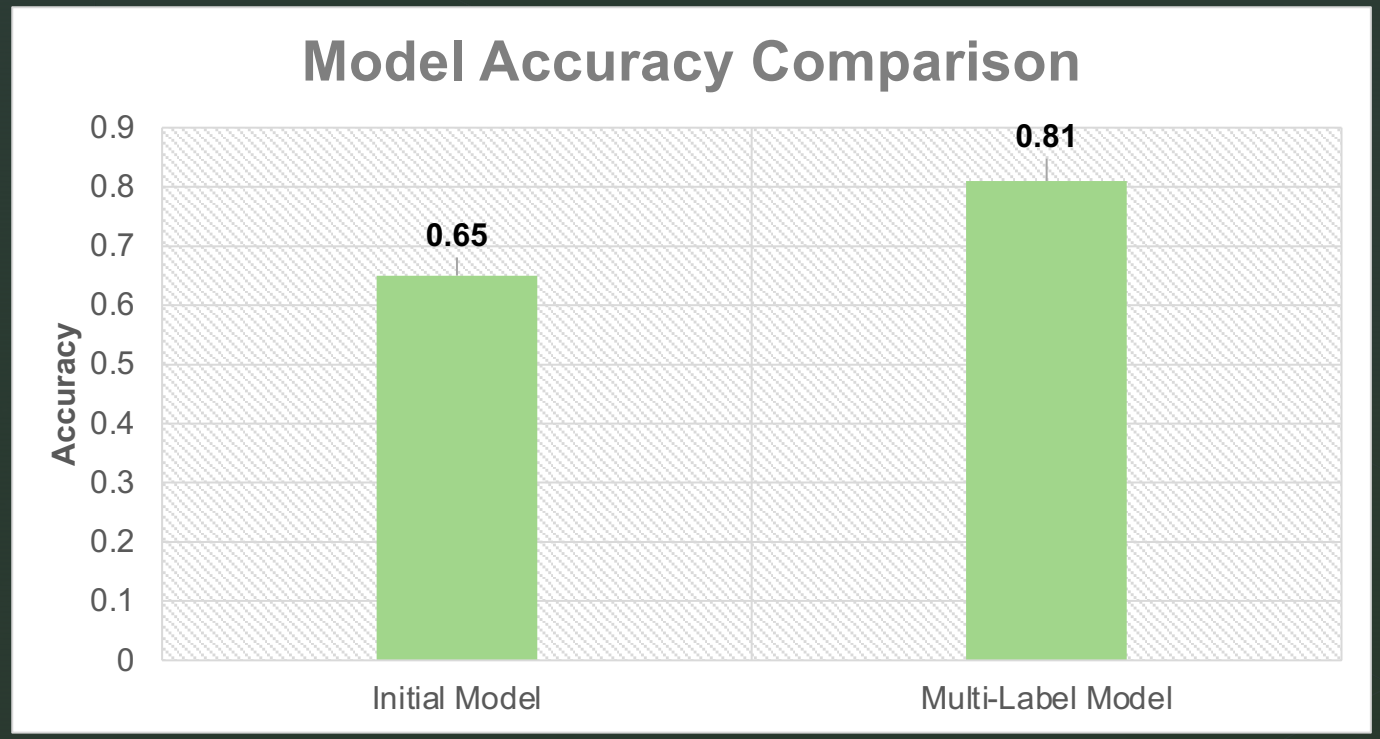
Classification of IT Material Weaknesses

- Categories derived from previous studies in IT Auditing

data processing integrity	system access and security	system structure and usage
Definition: the extent to which data is correct and reliable	Definition: the extent to which: data is available or easily and quickly retrievable; and access to data is restricted appropriately to maintain its security	Definition: the extent to which data is: easily comprehended; presented in the same format
weak IT **monitoring**	logical **access** issues	lack of system documentation, policies, procedures
inadequate system to support business processes (**manually** intense)	security issues	disparate (non-integrated systems)
weak application controls	(business user) **segregation of duties**	insufficient **training** on system
development and maintenance, **program change control** issues	inadequate records and storage retention	weak IT information and **communication**
weak end-user computing controls	lack of disaster recovery plan for systems	decentralized systems
weak general controls	IS/IT **personnel access not properly segregated**	
weak IT control environment		
inadequate IT/IS support staff		
weak IT control activities		
relying on system of others where controls not verified		
weak IT risk assessment		
lack of IS/IT controls		
data integrity issues		

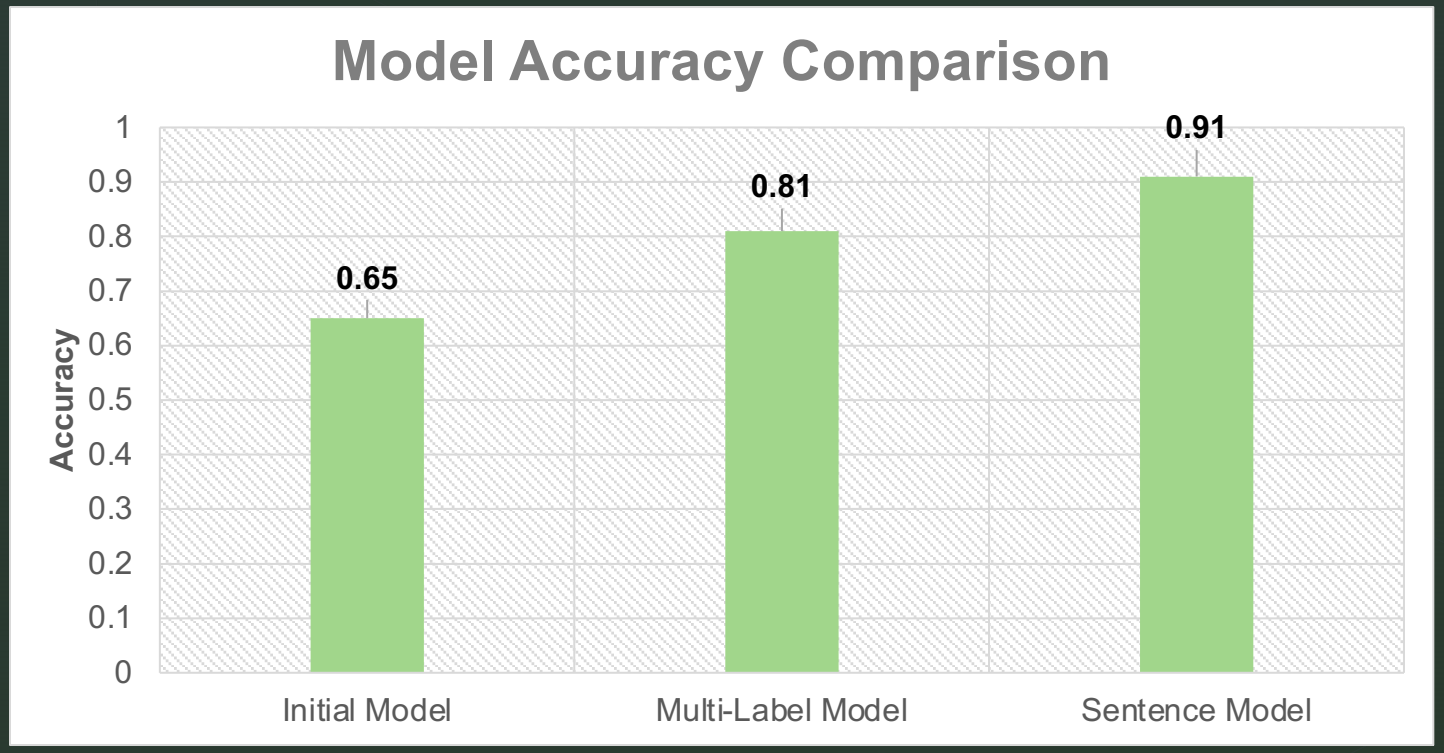
Weaknesses can be Belong to Multiple Categories

- Initial classification models suffered from classification errors.
- Observed that the extracted text could be categorized into multiple categories.
- To address classification errors, Multi-label classification was performed. Model performance improved.



A Better Way to Classify: Sentence Classification

- Divide extracted information into sentences and classify each sentence into a category.
- Sentence classification much more successful.



Recommendation: Implement NLP For Information Extraction

Successfully created a tool which can extract relevant information from a collection of documents.

The tool can be customized to fit an organizations needs.

The tool provides business value by saving time & resources, extracting complete and accurate information, with the possibility to discovering hidden / hard to find information.