

Weekly reports for IARCS Internship

Wong Chu Feng

Jan - May 2024

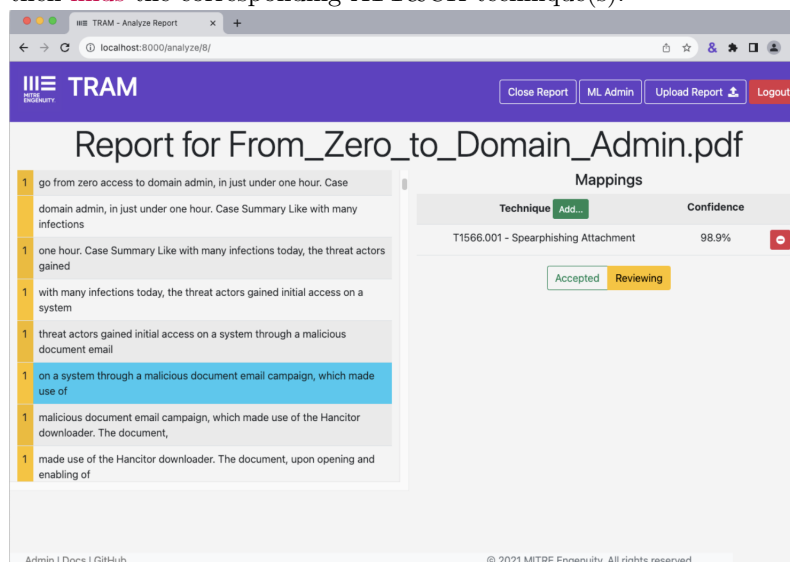
Contents

Week 1: Research on STRIDE and Mitre ATT&CK	2
Week 2: Building the model	4
Week 3: Finding better models	7
Week 4: Refining the corpus and model	11

Week 1: Research on STRIDE and Mitre ATT&CK

Pointers:

- Read up on STRIDE framework, MITRE ATT&CK and OpenAI's API.
- Briefing on internship, phase 1 and 2.
 1. **Attack technique classifier**
 - Try and evaluate existing ML/NLP techniques to classify ATT&CK techniques into STRIDE categories
 - Evaluate using LLM to classify ATT&CK techniques into STRIDE categories
 - Customize LLM to classify ATT&CK technique using other threat model categories (e.g., RAPIDS)
 2. **Architecture diagram analyzer**
 - Identify trust zones in architecture diagram based on subnet information
 - Identify unrealistic hand-drawn architecture diagrams
- Studied code from current stride (supervised) **model classifier**.
- Understood the supervised model's code. Looking at the *raw_capec_data.xlsx* to try clustering methods to create an unsupervised model. The trained supervised model might be useful in feature extraction to train the unsupervised model. What features exactly? Or in transfer learning? Still not too sure.
- Digressing a bit, started researching for similar papers on mapping ATT&CK to STRIDE.
- Discovered **TRAM LLM** → *Possible to use with the questionnaire from GovTech as an input.*
 - identifies keywords, capable of contextual understanding
 - predicts the presence of TTPs (Tactics, Techniques, Procedures) in the text
 - then **finds** the corresponding ATT&CK technique(s).



- **Identification of Attack Paths Using Kill Chain and Attack Graphs** → *Needs a separate function to create this graph. Seemingly not helpful.*

TABLE III
ATTACK GRAPH VERTICES

ID	Description
1	External actor.
2	Email address of an employee was published on a website.
3	T1594 – Search Victim-Owned Websites.
4	The attacker knows that the email address exists.
5	Sender reputation analysis was not accomplished.
6	T1566.001 – Spearphishing Attachment.
7	Authentication of sending an email was violated.
8	The employee can click on the attachment.
9	The employee has a user account on a personal computer.
10	Training of users was not accomplished (countermeasure).
11	T1204.002 – User execution of malicious file.
12	Authentication of opening file action was violated.
13	Microsoft Office opens files.
14	Microsoft Office is installed on the personal computer.
15	Microsoft Office 2016 contains CVE-2017-0262 vulnerability.
16	T1203 – Exploitation for Client Execution.
17	The attacker violated the system’s authorization (user rights).
18	Microsoft Windows 8.1 is installed on the personal computer.
19	Microsoft Windows 8.1 contains CVE-2017-0263 vulnerability.
20	Software is not regularly updated.
21	T1068 – Exploitation for Privilege Escalation.
22	The attacker violated the system’s authorization (admin rights).
23	T1485 – Data Destruction.
24	Data backup was not accomplished (countermeasure).
25	T1486 – Data Encrypted for Impact.
26	Integrity of a sensitive file was violated.

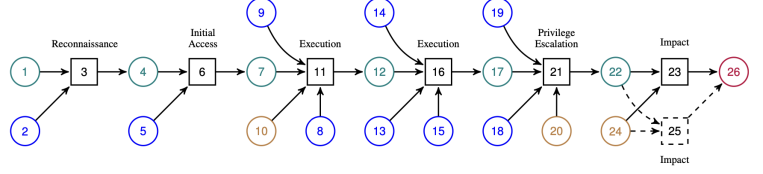


Fig. 3. Example kill chain attack graph for the validation use case. Vertices correspond to attacker’s level of control over assets (green), asset properties (blue), countermeasures (brown), attack techniques (black), and an attack goal (dark red). Vertices are described in Table III.

- Thinking of either a standalone unsupervised model or a **hybrid** one.
- Summary: Make use of the current supervised model that maps Mitre Att&cks using keywords to STRIDE. Then we can spatially obtain the clusters of data each corresponding to a category in STRIDE. We can use this as the baseline to train the unsupervised model. Each similarity of *category from GovTech* will be weighted against STRIDE, thereby allowing the mapping of STRIDE to the *GovTech framework*.
- Consolidate ideas:
 - Use unsupervised model.
 - Use hybrid model. Use the current supervised model to extract features (if any) and pass it to the unsupervised model to train for the *GovTech framework*.
 - Find the similarities between STRIDE and the *GovTech framework*, give each category some keywords and use DL to find its weights. Then use the weights to train the unsupervised model.
- Attempt to change the current supervised model to an unsupervised one.

Week 2: Building the model

Attempted at using *Tokenizer* for feature extraction. Yields 0.469 accuracy. Should be due to the corpus being too small for the model to properly learn the context of the words.

```
1 tokenizer = Tokenizer(oov_token='<OOV>')
2 tokenizer.fit_on_texts(df_train['NameDesc'])
3 X_train = tokenizer.texts_to_sequences(df_train['NameDesc'])
4 X_test = tokenizer.texts_to_sequences(df_test['NameDesc'])
5 X_val = tokenizer.texts_to_sequences(df_dev['NameDesc'])
6
7 x = [X_train, X_test, X_val]
8 max_length = 0
9 for _ in x:
10     max_l = max([len(seq) for seq in _])
11     max_length = max(max_length, max_l)
12
13 X_train_padded = pad_sequences(X_train, maxlen=max_length, padding='post')
14 X_test_padded = pad_sequences(X_test, maxlen=max_length, padding='post')
15 X_val_padded = pad_sequences(X_val, maxlen=max_length, padding='post')
16
17 y_train = df_train['STRIDE'].values
18 y_test = df_test['STRIDE'].values
19 y_val = df_dev['STRIDE'].values
```

Second attempt using *TfidfVectorizer* for feature extraction. This model performs better because TfidfVectorizer works better with a small corpus as it takes into account the frequency of the words appearing which might indicate keywords of a STRIDE category.

```
1 tfidf_vectorizer = TfidfVectorizer()
2
3 X_train_tfidf = tfidf_vectorizer.fit_transform(df_train['NameDesc']).toarray()
4 X_test_tfidf = tfidf_vectorizer.transform(df_test['NameDesc']).toarray()
5 X_val_tfidf = tfidf_vectorizer.transform(df_dev['NameDesc']).toarray()
6
7 y_train = df_train['STRIDE'].values
8 y_test = df_test['STRIDE'].values
9 y_val = df_dev['STRIDE'].values
```

Then I perform hyperparameter tuning on the model.:

```
1 dropout_rates = [0.2, 0.3, 0.4, 0.5]
2 activations_list = ['relu', 'leaky_relu', 'elu', 'tanh']
3 num_neurons = [32, 64, 128, 256]
4 opt_lr = [1e-2, 1e-3, 1e-4]
5 L2_lr = [1e-2, 1e-3, 1e-4]
6 best_params = None
7 best_val_acc = 0
8
9 hyperparam_combi = itertools.product(dropout_rates, num_neurons, activations_list,
10                                     opt_lr, L2_lr)
11
12 for dr, nn, al, olr, l2lr in hyperparam_combi:
13     modelTest = tf.keras.Sequential([
14         tf.keras.layers.Input(shape=(vocab_size,)),
15         tf.keras.layers.Dense(nn*2, activation=al),
16         tf.keras.layers.Dropout(dr),
17         tf.keras.layers.BatchNormalization(),
18         tf.keras.layers.Dense(nn, activation=al),
19         tf.keras.layers.Dropout(dr),
20         tf.keras.layers.BatchNormalization(),
21         tf.keras.layers.Dense(nn//2, activation=al),
22         tf.keras.layers.Dense(num_classes, kernel_regularizer=tf.keras.regularizers.L2(l2=1e-2), activation='softmax')
23     ])
24
25     optimizer = tf.keras.optimizers.legacy.Adam(olr)
26     modelTest.compile(optimizer=optimizer, loss="sparse_categorical_crossentropy",
27                     metrics=['accuracy'])
```

```

26
27     early_stop = EarlyStopping(
28         monitor="val_loss",
29         patience=5,
30         verbose=0,
31         restore_best_weights=True
32     )
33     histTest = modelTest.fit(
34         X_train_tfidf, y_train,
35         batch_size=16,
36         epochs=num_epochs,
37         validation_data=(X_val_tfidf, y_val),
38         verbose=0,
39         callbacks=[early_stop,]
40     )
41
42     val_acc = max(histTest.history['val-accuracy'])
43     # print(f"Dropout: {dr}, Activation: {al}, Hidden Units: {nn}, L2 Reg: {l2lr}, LR: {
44         olr}, Best Val Acc: {val_acc}\n=====")
45     if val_acc > best_val_acc:
46         best_val_acc = val_acc
47         best_params = (dr, nn, al, olr, l2lr)

```

Final Best Hyperparameters:

Dropout: 0.2,

Activation: *elu*,

Hidden Units: 32,

L2 Reg: 0.001,

LR: 0.0001,

Best Val Acc: 0.918367326259613

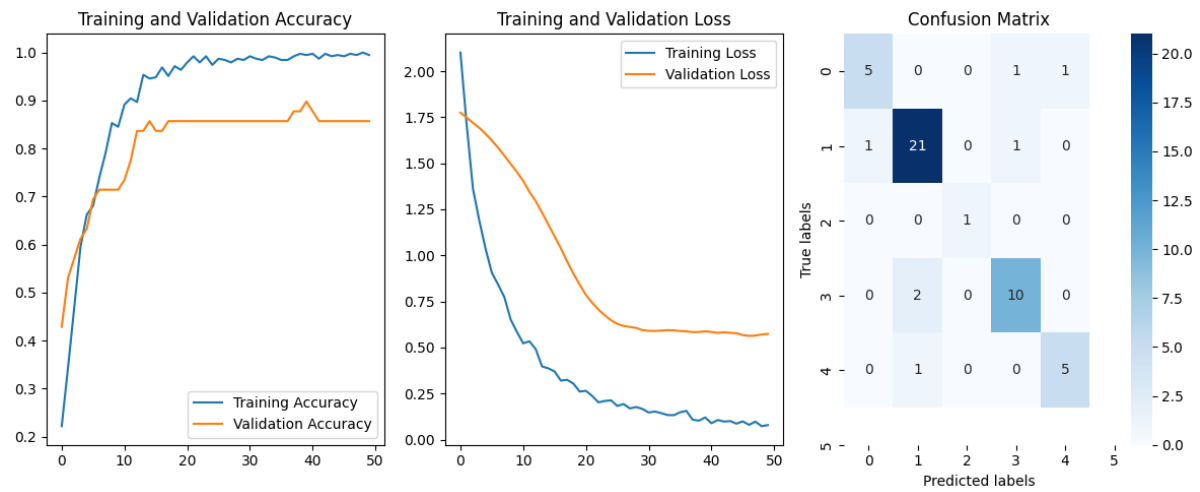
Final model used:

```

1 hidden_units = 32
2 num_classes = 6
3 batch_size = 16
4 num_epochs = 50
5 classes = [0,1,2,3,4,5]
6 vocab_size = X_train_tfidf.shape[1]
7 optimizer = tf.keras.optimizers.legacy.Adam(1e-4)
8
9 model3 = tf.keras.Sequential([
10     tf.keras.layers.Input(shape=(vocab_size,)),
11     tf.keras.layers.Dense(hidden_units*2, activation='elu'),
12     tf.keras.layers.Dropout(.2),
13     tf.keras.layers.BatchNormalization(),
14     tf.keras.layers.Dense(hidden_units, activation='elu'),
15     tf.keras.layers.Dropout(.2),
16     tf.keras.layers.BatchNormalization(),
17     tf.keras.layers.Dense(hidden_units//2, activation='elu'),
18     tf.keras.layers.Dense(num_classes, kernel_regularizer=tf.keras.regularizers.L2(1e-3), activation='softmax')
19 ])
20
21 model3.compile(optimizer=optimizer, loss="sparse_categorical_crossentropy", metrics=['
    accuracy'])
22 model3.summary()
23 # plot_model(model3, show_shapes=True, show_layer_names=True)

```

Results: Plotted a confusion matrix to evaluate the model. It yields an accuracy of 0.857 where the diagonal entries are the true labels predicted.



Summary:

- Try fuzzy clustering on the dataset since GMM and DBSCAN proves to be not too helpful.
- Attempt pulling out more specific keywords and predict the STRIDE category
- Examine using one model for each of STRIDE category

Week 3: Finding better models

Note: I discovered an error in the code for the previous model shown in week 2 where the promising results were ultimately spurious.

This is a more aggressive text preprocessing on the corpus. I first conduct a preliminary analysis of the word frequencies and trivially remove unimportant words.

```
1 def basic_processing(df):
2     words_to_remove = ["e.g.", "code", "may", "attack", "system", "adversary", "Adversaries"]
3     for word in words_to_remove:
4         df['NameDesc'] = df['NameDesc'].apply(lambda x: x.replace(word, ''))
5     for word in words_to_remove:
6         df['NameDesc'] = df['NameDesc'].apply(lambda x: re.sub(r'\b' + re.escape(word) + r'\b', '', x))
7
8     # df['NameDesc'] = df['NameDesc'].str.replace(r"\b(" + "".join(words_to_remove) + ")\b", "", regex=True)
9     df['NameDesc'] = df['NameDesc'].str.replace("<br><br>", "", regex=True)
10    df['NameDesc'] = df['NameDesc'].str.replace("\((Citation:.*?)\)", "", regex=True)
11    df['NameDesc'] = df['NameDesc'].str.replace("http\S+", "", regex=True)
12    df['NameDesc'] = df['NameDesc'].str.replace("-+", "-", regex=True)
13    df['NameDesc'] = df['NameDesc'].str.replace("[^A-Za-z]", "", regex=True)
14    return df
15
16 def rm_stopwords(df):
17     stop_words = set(stopwords.words('english'))
18     df['NameDesc'] = df['NameDesc'].apply(lambda x: [word for word in x if word not in stop_words])
19     print(f"Removed stopwords:\n-{df.head(3).NameDesc}\n")
20     return df
21
22 def lemmatize(df):
23     lemmatizer = WordNetLemmatizer()
24     def lemmatize_tokens(tokens):
25         def get_wordnet_pos(word):
26             tag = nltk.pos_tag([word])[0][1][0].upper()
27             tag_dict = {"J": wordnet.ADJ,
28                        "N": wordnet.NOUN,
29                        "V": wordnet.VERB,
30                        "R": wordnet.ADV}
31             return tag_dict.get(tag, wordnet.NOUN)
32         lemmas = [lemmatizer.lemmatize(token, get_wordnet_pos(token)) for token in tokens]
33         return lemmas
34     df['NameDesc'] = df['NameDesc'].apply(lambda x: lemmatize_tokens(x))
35     print(f"Lemmatized words:\n-{df.head(3).NameDesc}")
36     return df
37
38 def text_preprocessing(df):
39     basic_processing(df)
40     df['NameDesc'] = df['NameDesc'].apply(lambda x: word_tokenize(x))
41     rm_stopwords(df)
42     lemmatize(df)
43     print("=====")
44     return df
```

Table 1: Extracting keywords

Dataset	Original text	Stemmed & Lemmatized text
df_train[0]	Exfiltration to Cloud Storage Adversaries may exfiltrate data to a cloud storage service rather than over their primary command and control channel. Cloud storage services allow for the storage, edit, and retrieval of data from a remote cloud storage server over the Internet. Examples of cloud storage services include Dropbox and Google Docs. Exfiltration to these cloud storage services can provide a significant amount of cover to the adversary if hosts within the network are already communicating with the service. Analyze network data for uncommon data flows (e.g., a client sending significantly more data than it receives from a server) to known cloud storage services. Processes utilizing the network that do not normally have network communication or have never been seen before are suspicious. User behavior monitoring may help to detect abnormal patterns of activity.	[Exfiltration, Cloud, Storage, exfiltrate, data, cloud, storage, service, rather, primary, command, control, channel, Cloud, storage, service, allow, storage, edit, retrieval, data, remote, cloud, storage, server, Internet, Examples, cloud, storage, service, include, Dropbox, Google, Docs, Exfiltration, cloud, storage, service, provide, significant, amount, cover, host, within, network, already, communicate, service, Analyze, network, data, uncommon, data, flow, client, send, significantly, data, receives, server, know, cloud, storage, service, Processes, utilize, network, normally, network, communication, never, see, suspicious, User, behavior, monitoring, help, detect, abnormal, pattern, activity]
df_train[2]	Runtime Data Manipulation Adversaries may modify systems in order to manipulate the data as it is accessed and displayed to an end user, thus threatening the integrity of the data.(Citation: FireEye APT38 Oct 2018)(Citation: DOJ Lazarus Sony 2018) By manipulating runtime data, adversaries may attempt to affect a business process, organizational understanding, and decision making. Adversaries may alter application binaries used to display data in order to cause runtime manipulations. Adversaries may also conduct [Change Default File Association](https://attack.mitre.org/techniques/T1546/001) and [Masquerading](https://attack.mitre.org/techniques/T1036) to cause a similar effect. The type of modification and the impact it will have depends on the target application and process as well as the goals and objectives of the adversary. For complex systems, an adversary would likely need special expertise and possibly access to specialized software related to the system that would typically be gained through a prolonged information gathering campaign in order to have the desired impact. Inspect important application binary file hashes, locations, and modifications for suspicious/unexpected values.	[Runtime, Data, Manipulation, modify, order, manipulate, data, access, displayed, end, user, thus, threaten, integrity, data, By, manipulate, runtime, data, adversary, attempt, affect, business, process, organizational, understand, decision, make, alter, application, binary, use, display, data, order, cause, runtime, manipulation, also, conduct, Change, Default, File, Association, Masquerading, cause, similar, effect, The, type, modification, impact, depends, target, application, process, well, goal, objective, For, complex, would, likely, need, special, expertise, possibly, access, specialized, software, related, would, typically, gain, prolong, information, gathering, campaign, order, desire, impact, Inspect, important, application, binary, file, hash, location, modification, suspicious, unexpected, value]

Visualising the word frequencies in the corpus:

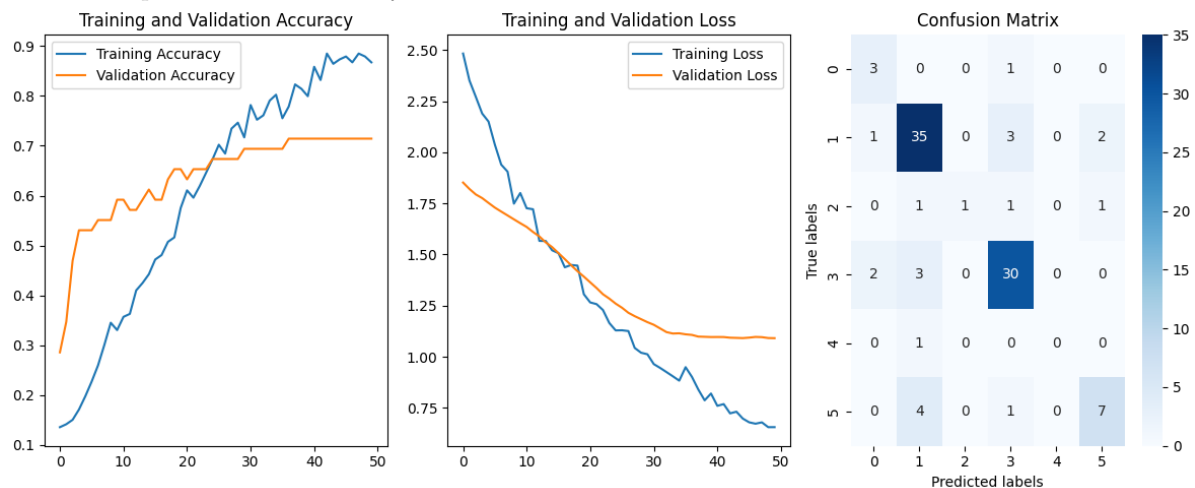
Using TfidfVectorizer and after hyperparameter tuning, the model to use is as follows.

```

1 hidden_units = 32
2 batch_size = 16
3 num_epochs = 50
4 num_classes = 6
5 classes = [0,1,2,3,4,5]
6 vocab_size = X_train_tfidf.shape[1]
7 optimizer = tf.keras.optimizers.legacy.Adam(1e-4)
8
9 model4 = tf.keras.Sequential([
10     tf.keras.layers.Input(shape=(vocab_size,)),
11     tf.keras.layers.Dense(hidden_units*2, activation='elu'),
12     tf.keras.layers.Dropout(.5),
13     tf.keras.layers.BatchNormalization(),
14     tf.keras.layers.Dense(hidden_units, activation='elu'),
15     tf.keras.layers.Dropout(.5),
16     tf.keras.layers.BatchNormalization(),
17     tf.keras.layers.Dense(hidden_units//2, activation='elu'),
18     tf.keras.layers.Dense(num_classes, kernel_regularizer=tf.keras.regularizers.L2(1e-2), activation='softmax')
19 ])
20
21 model4.compile(optimizer=optimizer, loss="sparse_categorical_crossentropy", metrics=['
22     accuracy'])
23 model4.summary()
# plot_model(model4, show_shapes=True, show_layer_names=True)

```

This model predicts with accuracy of 0.79.

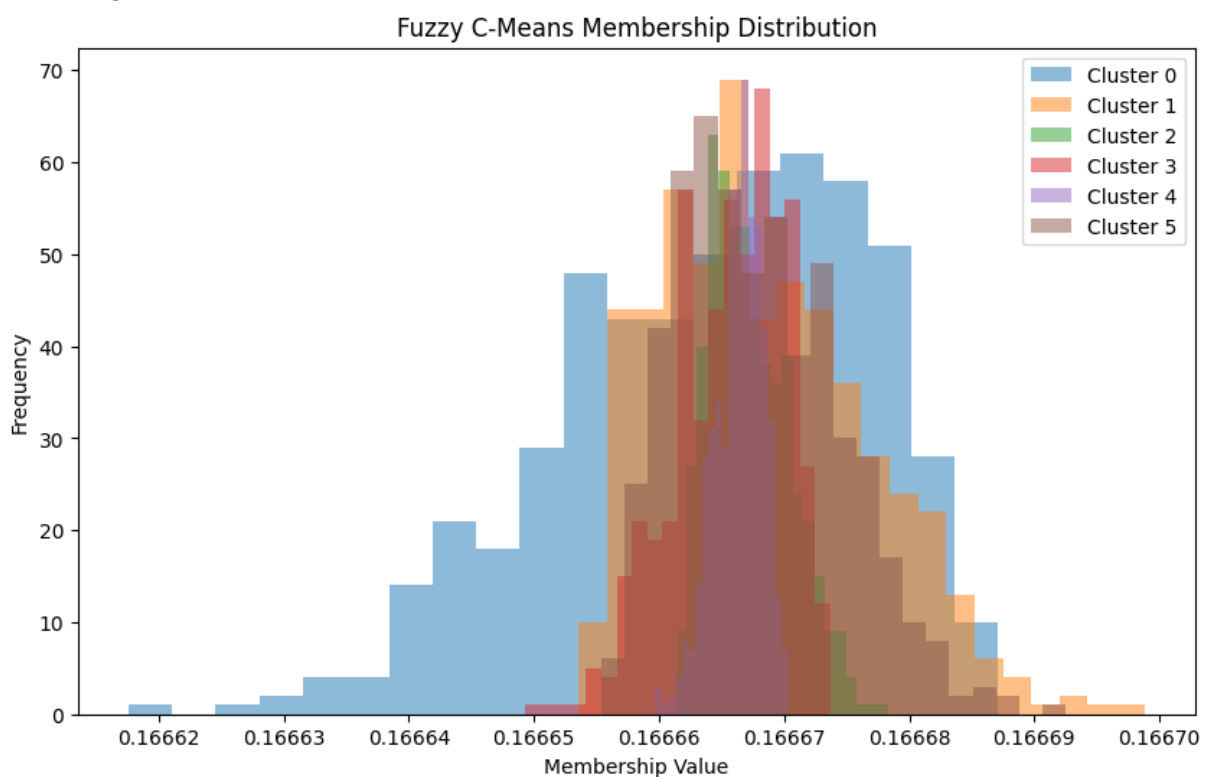


Attempt on Fuzzy Clustering

Table 2: Extracting keywords

Cluster	Unique STRIDE values
0	['000001' '010000' '000100' '000010' '110000' '010001' '100000' '101000' '100100' '011000']
1	['000010' '010000' '000001' '000100' '100000' '010010' '011000' '100001' '100100' '101000']
2	['000001' '010000' '000100']
3	['011000']
4	['000001' '010000' '010100' '000100' '100000']
5	['010001' '000001' '010000' '011000' '000010' '000100' '100100' '100000']

Visualising the clusters:



Moving forward, there are several improvements to be made.

- Instead of selecting keywords from the corpus, it can be constructed by hand using the definition of each categories of STRIDE, and subsequently of RAPIDS.
- Filtering of keywords can include more manual methods to further reduce the keyword corpus.
- Visualise and examine the frequencies for some keywords and manually remove some, then we will be left with lesser keywords that the model will be used to train on, potentially increasing the accuracy of classification.

Week 4: Refining the corpus and model

Improvement 1: Removed duplicate words

Table 3: Extracting keywords v2.0

Dataset	Original text	Processed text
df_train[0]	<p>Business Relationships Adversaries may gather information about the victim’s business relationships that can be used during targeting. Information about an organization’s business relationships may include a variety of details, including second or third-party organizations/domains (ex: managed service providers, contractors, etc.) that have connected (and potentially elevated) network access. This information may also reveal supply chains and shipment paths for the victim’s hardware and software resources.

Adversaries may gather this information in various ways, such as direct elicitation via [Phishing for Information](https://attack.mitre.org/techniques/T1598). Information about business relationships may also be exposed to adversaries via online or other accessible data sets (ex: [Social Media](https://attack.mitre.org/techniques/T1593/001) or [Search Victim-Owned Websites](https://attack.mitre.org/techniques/T1594)).(Citation: ThreatPost Broadvoice Leak) Gathering this information may reveal opportunities for other forms of reconnaissance (ex: [Phishing for Information](https://attack.mitre.org/techniques/T1598) or [Search Open Websites/Domains](https://attack.mitre.org/techniques/T1593)), establishing operational resources (ex: [Establish Accounts](https://attack.mitre.org/techniques/T1585) or [Compromise Accounts](https://attack.mitre.org/techniques/T1586)), and/or initial access (ex: [Supply Chain Compromise](https://attack.mitre.org/techniques/T1195), [Drive-by Compromise](https://attack.mitre.org/techniques/T1189), or [Trusted Relationship](https://attack.mitre.org/techniques/T1199)).

Much of this activity may have a very high occurrence and associated false positive rate, as well as potentially taking place outside the visibility of the target organization, making detection difficult for defenders.

Detection efforts may be focused on related stages of the adversary lifecycle, such as during Initial Access.</p>	<p>['websites', 'associate', 'contractor', 'supply', 'path', 'online', 'variety', 'operational', 'this', 'phishing', 'gathering', 'via', 'search', 'use', 'adversary', 'open', 'take', 'initial', 'drive', 'victim', 'include', 'etc', 'manage', 'potentially', 'trusted', 'access', 'connect', 'target', 'resource', 'elicitation', 'hardware', 'organization', 'direct', 'place', 'rate', 'elevate', 'network', 'service', 'well', 'domains', 'second', 'activity', 'gather', 'way', 'establish', 'stage', 'provider', 'compromise', 'data', 'defender', 'relationship', 'false', 'chain', 'reconnaissance', 'accounts', 'media', 'social', 'relationships', 'outside', 'information', 'effort', 'domain', 'detail', 'related', 'business', 'software', 'various', 'also', 'opportunity', 'lifecycle', 'ex', 'set', 'owned', 'positive', 'detection', 'third', 'accessible', 'party', 'high', 'difficult', 'occurrence', 'make', 'shipment', 'visibility', 'focus', 'reveal', 'expose', 'form']</p>
df_train[2]	<p>Symmetric Cryptography Adversaries may employ a known symmetric encryption algorithm to conceal command and control traffic rather than relying on any inherent protections provided by a communication protocol. Symmetric encryption algorithms use the same key for plaintext encryption and ciphertext de-</p>	<p>['content', 'sample', 'packet', 'know', 'obtain', 'data', 'flow', 'provide', 'with', 'uncommon', 'control', 'utilize', 'traffic', 'port', 'key', 'general', 'detect', 'de', 'algorithm', 'see', 'send', 'rc', 'follow', 'decryption', 'suspicious', 'employ', 'protection', 'processes', 'plaintext', 'cryptograph', 'malware', 'ciphertext', 'inherent']</p>

Figuring out keywords using Microsoft's definition of STRIDE

S: Involves illegally accessing and then using another user's authentication information, such as username and password

T: Involves the malicious modification of data. Examples include unauthorized changes made to persistent data, such as that held in a database, and the alteration of data as it flows between two computers over an open network, such as the Internet

R: Associated with users who deny performing an action without other parties having any way to prove otherwise—for example, a user performs an illegal operation in a system that lacks the ability to trace the prohibited operations. Non-Repudiation refers to the ability of a system to counter repudiation threats. For example, a user who purchases an item might have to sign for the item upon receipt. The vendor can then use the signed receipt as evidence that the user did receive the package

I: Involves the exposure of information to individuals who are not supposed to have access to it—for example, the ability of users to read a file that they were not granted access to, or the ability of an intruder to read data in transit between two computers

D: Denial of service (DoS) attacks deny service to valid users—for example, by making a Web server temporarily unavailable or unusable. You must protect against certain types of DoS threats simply to improve system availability and reliability

E: An unprivileged user gains privileged access and thereby has sufficient access to compromise or destroy the entire system. Elevation of privilege threats include those situations in which an attacker has effectively penetrated all system defenses and become part of the trusted system itself, a dangerous situation indeed

Create keywords from the definitions

S: ['authenticate', 'username', 'password', 'access']

T: ['modify', 'persistent', 'database', 'alter', 'open', 'network', 'internet']

R: ['deny', 'action', 'prove', 'non-repudiation', 'item', 'sign', 'receipt', 'receive', 'evidence', 'package', 'untrace',]

I: ['exposure', 'individual', 'access', 'file', 'granted', 'intruder', 'transit']

D: ['denial', 'service', 'dos', 'web', 'server', 'unavailable', 'unusable', 'system', 'available', 'reliable']

E: ['unprivileged', 'privileged', 'access', 'compromise', 'entire', 'system', 'elevation', 'penetrate', 'defenses', 'untrusted', 'trusted']