Towards Establishing the Criteria to be Considered in Applying Natural Language Processing to Forensic Investigations on Emails

Anonymous[1][0000-1111-2222-3333] and Anonymous[2][1111-2222-3333-4444]

[1] Anonymous University, Princeton NJ 08544, USA
[2] Springer Heidelberg, Tiergartenstr. 17, 69121 Heidelberg, Germany
anonymous@springer.com

Abstract. Digital forensic investigators often face a large number of emails seized in an investigation. Processing of the emails to find evidence can be supplemented with natural language processing techniques. The academic application of natural language processing may not make the task of the investigator easier, unless the techniques are applied in a manner such that it matches the investigative process. This paper investigates the main considerations and requirements to find a solution for applying natural language processing in digital forensics investigations. This paper ends of by presenting a general architecture of the main elements required in an application of natural language processing in emails.

Keywords: Digital Forensic Investigations, Email, Natural Language Processing.

1 Introduction

Computer forensic investigations involve the seizure and analysis of large volumes of data, including emails, documents, computer logs, partially deleted data [1]. Computer forensic analysts go through various stages during the investigation. ISO 27043 [2] breaks the forensics process down into classes, of which post-incident classes are initialization, acquisitive and investigative. The activities described during these phases are respond, identify, collect, acquire and preserve, understand, report and close. For the purposes of this paper it is assumed that the initialization and acquisitive phases is already performed, and the focus is on the investigative part of the process. The associated activities in ISO 27043 are therefore understand, report and close. It is during the understanding phase that, dealing with large volumes of documents, a digital forensic investigator faces an enormous task. Various Natural Language Processing (NLP) techniques can potentially be used during this stage to assist the investigator in sifting through the data. Simple keyword search is the most basic form, but requires the investigator to have prior knowledge of the subject matter he/she is investigating, knowing a piori what he/she is looking for. This is not always the case as investigators often deal with subject matter they have not seen before [1] [3]. Simple keyword search does

not reveal other aspects that is important in the investigative process, e.g. timelines and spatial relationships, multi-theory development [4].

To supplement the investigative process and option is to use NLP techniques to produce summaries of texts. One such NLP technique is topic analysis. The aim in topic analysis is to analyze text and group sentences which are related in topic together [5]. This in itself may be useful, but large volumes of text will contain many topics, and although prior knowledge of keywords is not required from the investigator, it may still produce an overwhelming number of topics and keywords to the investigator. Text summarization is a next option, but similarly may leave the investigator overwhelmed.

Forensic investigations are expensive and generally charged on at hourly rates of the investigators involved in analyzing the data. There are also time constraints as the outcome of the investigation needs to be used in an outcome e.g. prosecution. Finding an optimal tool to assist in the investigative process which optimize the time spent by the investigator, especially in the initial phases of the investigation is not as simple as it may seem.

For the purposes of this paper, the focus is narrowed to the analysis of emails. Emails lend themselves to a natural spatial relationship and time ordering. Existing digital forensic tools extract entity relation diagrams to determine who communicates with whom by extracting the information from the email header data [1]. Emails are automatically ordered by timestamp. Preprocessing of attachments is automated including Optical Character Recognition (OCR), so that plaintext versions of scanned attachments are available. Although subject lines are available, it is still challenging to find information in emails quickly. The subject line and email body may not always be related, or important information in the email body not covered by the subject line in the header may be of interest to the investigator.

By studying available NLP techniques and the investigative process limited to the processing of emails, this paper attempts to answer the following questions:

1. What are the main investigative and technical considerations when relying on NLP techniques in forensic investigations?
2. What are the basic requirements for an NLP solution to be applied to emails in digital forensic analysis?
3. What elements would a general NLP architecture consist of to process emails into useful summaries for the forensic investigator?

The rest of this paper is structured as follow. Section 2 introduces some background to NLP solutions and the principles they rely on to work successfully. Additionally, section 2 introduces some basic of investigative processes as presented in [4]. Section 3 derives the technical, investigative and ethical considerations when applying NLP to emails, as well as some requirements to an NLP solution to be applied to emails. Section 4 proposes a general solution architecture to meet these requirements. Section 5 deals with future work while section 7 deals with related work. Section 7 summarize and concludes the paper.

2 Background

This section introduces the basic background required to understand the investigative process, and natural language processing. It is not meant to be a comprehensive introduction to the subject areas, and the interested reader can refer to referenced material for a more thorough introduction. Section 2.1 covers the principles applied in investigations as presented in [4]. Section 2.2 introduces modern natural language processing principles applied with machine learning techniques. A good introduction can be found at [5], [6] and [7], which this section is based on.

2.1 Investigation Principles

Experienced investigators develop their own unique approach or methodology to investigations. A general methodology is introduced to train new investigators. This methodology can be summarized by the steps given by the acronym STAIR [4]:

- Situation: Understanding the bigger picture in order to classify and prioritize a response.
- Tasks: Gathering, protecting and preserving evidence and information.
- Analysis: Examining the facts and physical evidence to develop theories.
- Investigate: Validating the facts through corroboration of witness accounts and forensic analysis.
- Results: Prioritizing and focusing on the results to guide the investigative process.

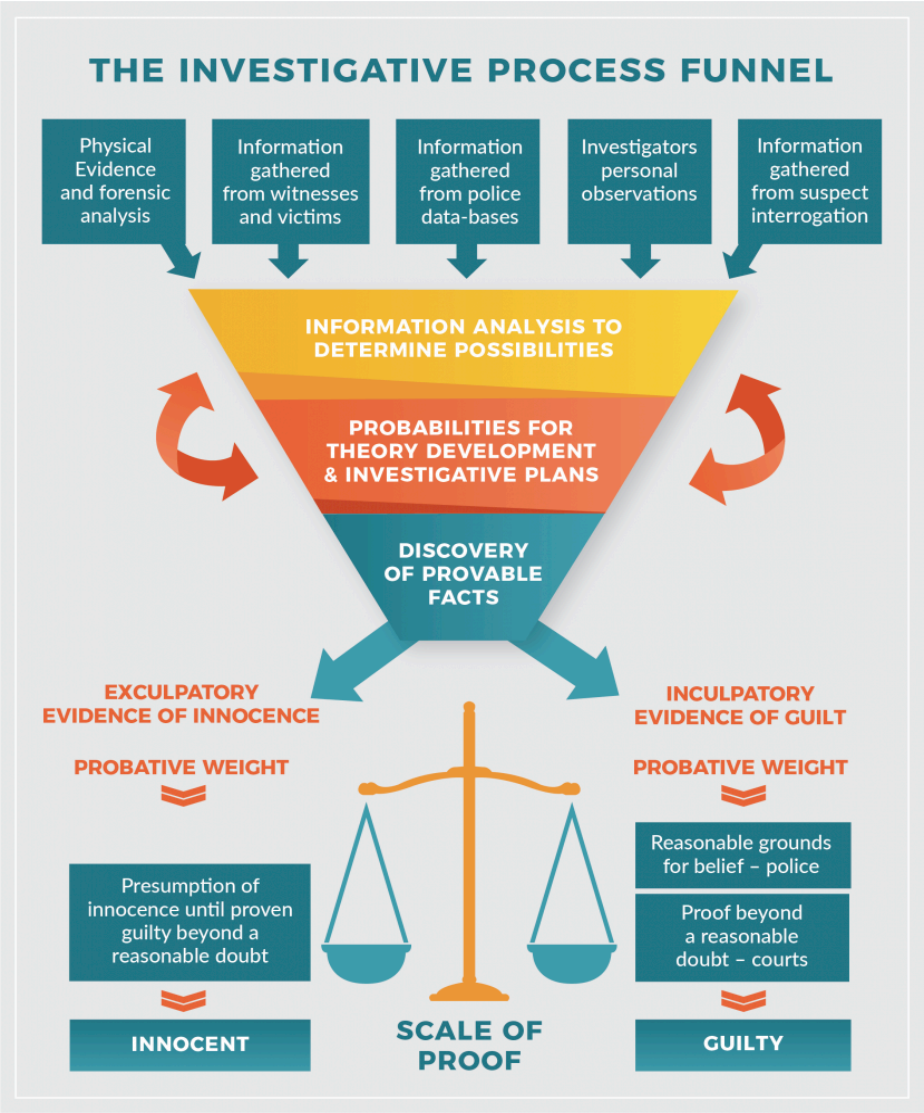
Analysis and investigation is an iterative process. Theories are developed in the analysis phase and tested in the investigation phase. With new insights, new theories are developed and tested again in the investigation phase. There is no explicit limit to the number of iterations that can take place. This is generally described as multi theory development.

Two types of investigative responses occur in general, namely the tactical response and the strategic response [4]. The tactical response investigation is urgent and requires a police officer to focus on assessing a tactical situation and saving lives. Once the threat is no longer there, and lives or property is not at stake, the investigation turns into a strategic investigation. During the strategic investigation phase, the evidence is collected, witness statements are taken, investigation actions are documented, and theories are developed. The strategic investigation phase is slow paced and deliberate [4]. The investigation process is best described by Figure 1 taken from [4].

The figure is largely self-explanatory, but the emphasis of the following aspects is important for this paper. The investigative processes make use of many sources of information at the top of the funnel, which is analyzed together. The analysis leads to multiple theory development, with probabilities associated with the theories. The theories lead to the discovery of provable facts which can either lead to confirmation of innocence or guilt. In fact, a common defense strategy is to find evidence, develop probable theories and present it to the court to cast doubt on the accuracy of the prosecution's theories [1].

For the purposes of this paper, the focus area is on digital forensic analysis of emails. It is represented by the left input into the investigative funnel in Figure 1. Any techniques applied during the digital forensic analysis must support both outcomes of proving innocence or guilt. Any digital forensic process must not be biased towards either one of the outcomes, or influence theories by an inherent bias in the tools.

Fig. 1.



Summary of the Investigative Process [4]

Notice the arrows between the information analysis phase and the probabilities for the theory development phase. This part of the investigation is iterative [4]. It implies that the investigator needs to revisit the analysis, conduct new analysis based on theories developed (and possibly) develop new analysis to test theories and their probabilities.

For the purposes of this paper, two more aspects covered in [4] regarding the investigative process is emphasized namely spatial relationships and timeline evidence. "Spatial relationships are circumstantial links that demonstrate connections between objects, events, or people" [4]. This can be used to link a suspect to a crime. "Timeline evidence is any item that demonstrates a time alignment of the suspect to the criminal event or the victim" [4].

Finally, the common mistakes investigators make are [4]:

1. Failing to identify and collect the available evidence and information.
2. Failing to effectively analyze the evidence and information collected to identify suspects and form reasonable grounds to take action.
3. Becoming too quickly focused on one suspect or one theory of events and ignoring evidence of other viable suspects or theories that should be considered.

2.2 Natural Language Processing

There are two approaches to Natural Language Processing, namely a linguistics approach or a statistical approach. Statistical approaches are popular in machine learning and is explored further in this section. This is not meant to be a through introduction to NLP. The most important aspects to keep in mind when constructing NLP systems to aid investigation are highlighted in this section. Specifically, the focus is on introducing only relevant concepts that assist the understanding of NLP applications in the context of utilizing it to assist in digital forensic investigators.

Some basic definitions are required. Sentences are built up from words. Adding a number of sentences together forms a topic that an author wants to communicate. A document is a collection of sentences with covering one or more topics. The term document is used in the abstract sense. For instance, an email, or a Tweet can be a document, and documents attached to an email are also referred to as documents. A corpus is a collection of related or unrelated documents. To aid understanding, this paper will only consider "whole" documents, i.e. from a computer forensics perspective, partially recovered documents by means of hard drive scraping are not considered.

A dictionary of a language is a reference of all words in the language. When dealing with a document, only a fraction of the words in the dictionary are used. For efficiency reasons a document must be represented as a vector space of numbers representing only the words that are contained in the corpus. The vector representations of the words capture features of the words in the documents. For example, a feature representation can be a unique number representation of a word and its overall frequency count in the document. This is referred to as a Bag of Words (BOW). The bag contains all the words used in the corpus, and their associated term frequency count.

The occurrence of a word is not the only measure of importance. For instance, the word "the" in the English language will have a high term frequency count, but that does not make it the most important word. To compensate for this, word importance ranking can be considered as Term Frequency, Inverse Document Frequency (TF-IDF). This technique allows an adjustment on the measure by taking the term frequency and multiplying it by the inverse of the term frequency in all documents in the corpus.

Other possible transformations of the vector space [6] include Latent Semantic Indexing (LSI), Random Projections (RP), Latent Dirichlet Allocation (LDA) and Hierarchical Dirichlet Process. A popular way to perform topic analysis on a corpus is through using LDA. The details is beyond the scope of this paper. It is sufficient to understand that it is a statistical technique applied to the corpus, finding clusters, and identifying words cluster as being related to the same topic. The number of clusters is guessed beforehand, and therefore can lead to words incorrectly being assigned to a topic.

One problem with a bag of words approach is that the context of related words in a sentence is not taken into account. Two models developed to overcome this, are Skipgram [8] and Continuous Bag of Words (CBOW) [9]. This involves designing and training Neural Networks that attempt to capture the statistics of a given word in a sentence, i.e. it takes into account surrounding words in the sentence. The training of such neural networks be performed on the corpus under investigation. Alternatively pre-trained neural networks on large corpora can be used to analyze a given corpus by using the pre-trained neural network as a predictor on the corpus. This forms the basis of more advanced techniques like topic analysis, sentiment analysis etc. For the purposes of this paper, the most important aspect to understand is that the core of these models relies on statistics.

Natural language processing like any statistical or machine learning problem, requires several decisions to be made. First and foremost is the choice of the model(s) to be applied. Both supervised and unsupervised machine learning techniques exist for NLP. Examples of NLP models are:

- Distinguish between classes of texts. A well-known example of class distinction in text is email filtering. Email filtering takes place by training machine learning algorithms through with labeled emails as either normal or spam. The machine learning algorithm thereby learns the features which identifies these two classes of texts.
- Perform sentiment analysis. Sentiment analysis can be seen as a more advanced form of text classes. Machine learning algorithms are trained on e.g. reviews written by consumers, including the rating the reviewer gave to a product. The review score is an indication of the reviewer's sentiment about the product. The machine learning algorithm attempts to identify the combination of words associated with the review score. When new text is presented to the machine learning algorithm, the algorithm attempts to predict the sentiment score based on the review text.
- Topic Modelling. Topic modelling, unlike the previous examples, is performed by unsupervised machine learning techniques i.e. texts are not pre-labeled manually. In topic modelling the machine learning algorithm performs statistical analysis of features on the texts and groups the texts into related classes called topics.

– Text summarization. Text summarization utilize a combination of techniques like topic modeling and text ranking to extract text, finding related topics, ranking the topics and then presenting a short summary of the document.

The list is not exhaustive. For the purpose of this paper the focus is on unsupervised techniques. With a suitable model(s) chosen, the data (emails) needs to go through a number of steps to obtain useable output from the machine learning model.

Data preprocessing: The text corpus needs to be extracted from the source documents. Text normalization needs to take place. Normalization includes harmonizing spelling and spell correction, detecting and harmonizing abbreviations, replacing inflections with base words and removing stop words like “is” and “are” which does not contribute to understanding. Other possible pre-processing steps to consider are identifying named entities, detecting sentence boundaries and possibly word sense disambiguation. The list goes on. The preprocessing ends with breaking sentences up into tokens, i.e. the words that will be trained, and finding an appropriate presentation for these words (like Bag of Words).

Model training: With an appropriately chosen model(s), the preprocessed text corpus is used to train the NLP model. During model training automated techniques determine the most suitable internal parameters for the model to use during predictions. Model training also involves input of parameters that guides the training process e.g. the number of training steps to be taken. Some parameters are fixed by the data scientist once off, or may optionally be set by the user, depending on the application.

Model prediction: The trained model is applied to a preprocessed text corpus to extract relevant information by means of prediction. In topic analysis for instance, the prediction will extract the relevant topics. In topic clustering, clusters will be predicted and extracted for presentation.

Presentation: Presentation of the prediction results can be simple. For example an email classification can be indicated as normal or spam, a Tweet can be indicated as hate speech, or normal speech. Presentation of the result can also be complex. For example the result of topic analysis grouped into wordclouds or text summaries. This is arguably the most critical step in making the model useful to an end user.

3 Considerations in Applying Topic Analysis to Emails

The preceding section introduced the most important aspects to consider regarding the investigative process and NLP. In this section the considerations that need to be taken into account when designing an NLP solution to aid digital forensic investigators are derived. The basic idea is that the NLP solution will have to produce a summary of the emails to the investigator to assist the investigator to pick the correct combination of emails to analyze and investigate further.

Section 3.1 deals with investigation and technical considerations. Section 3.2 derives requirements for an NLP model applied to emails in digital forensic investigations. Section 3.2 deals with ethical considerations when applying machine learning in a digital forensics setting.

Training time. There are three technical constraints involved industrializing solutions of this nature, namely processing power, available memory and time. Training time of neural networks can be long on large corpora. The training time should not be so long that the investigator resolves to manual processing. Although GPUs are popular options in reducing training time, they can become prohibitively expensive and in some cases the electricity cost can become a factor. Using only pre-trained neural network solutions like GLOVE is an option at the possible cost of accuracy on the desired corpus. Pre-trained neural network solutions may in some cases not be available on local languages.

Reducing text vs losing detail. The statistical nature of NLP solutions means that when the text in emails are reduced, some detail will be lost. The most important words or terms that the investigator may need to consider are not necessarily the ones with the highest frequency of occurrence. In fact, the word or phrase may even appear only once. The implication is that there should be some variability in summary, i.e. running the summary again can lead to a slightly different summary, or alternative taking less emails into account can improve the summary of the emails.

Multi theory development. The investigator needs to develop multiple theories during the investigation. Developing multiple theories is an attempt by the investigator to remain as objective as possible, and also to avoid "tunnel vision" by considering as many explanations of the circumstantial evidence as possible. During multi-theory development, the theories are assigned probabilities when investigated.

Text summarization on the other hand will provide a single view of the documents being summarized. The machine learning algorithms that produce the summaries are inherently biased by their statistical nature. The algorithms are also largely fixed, i.e. given the same corpus, the same text summaries will be produced. This may lead to two problems.

- The inherent bias may lead to the investigator developing tunnel vision due to the bias introduced by the models.
- The summarized text inflexibility makes it difficult to recognize multiple theories.

To aid multi-theory development, and to reduce the inherent bias caused by the underlying models, some variability must be built into the models to produce slight variations into the summaries produced. By successive runs the investigator is presented with different views on the same underlying text, aiding his ability to recognize other possible theories that he might want to explore from the original emails.

Creating timelines and spatial relationships. Assuming that email corpora timestamps are synchronized, a source of timelines already exists in the metadata to the emails. The NLP techniques presented in the section 2.2 does not make use of a time ordering. The overall model must be adapted to present summaries in a timeline, or even better, the model must be applied to take timelines into consideration. Spatial relationships in NLP starts with recognizing entities. Named entities may need have a higher weight in summary calculation and ranking. Named entities may not be the only spatial links available. For example, a physical item like an invoice, or an action like invoicing may carry more weight in the ranking. A deeper understanding is required regarding this aspect to improve NLP models.

Reporting and presentation. The output of the NLP is a form of reporting to the investigator. The form in which this report is presented, is important. The aim is that the summarized report must reduce the workload considerably to the investigator. The report must not replace the original email trail. It must be presented in such a way that the investigator can identify the trail of emails to extract from the evidence file and focus attention on the extracted emails. With emails a visual presentation of summaries may aid the presentation of summaries to the investigator considerably.

Reproducibility and interpretability. The investigator needs to document the investigation process, not only the results. The defense team may ask for documentation so that the investigation process can be scrutinized. Given the same data and process, the defense team therefore must be able to reproduce the same summaries from the emails if the documented investigation procedure is followed. If not, the defense may argue that the investigator may have been biased. An investigator must not be burdened by documenting each and every technical step he takes, taking his focus off the task of investigating. A full solution should therefore keep track of the investigator's actions that may lead to a different outcome. An important question that arise in investigations involving machine learning is interpretability [10]. Not only the investigator's bias is under scrutiny when interpreting the investigation results, but also the tools and techniques the investigator used. The NLP techniques must therefore be interpretable.

Measuring success. Developing an NLP model requires some type of measure on its success. A subjective measure is to provide a solution to experienced investigators and getting their feedback. On small datasets the basic model function can possibly be validated by a manual means. On large datasets the measure of success is difficult to assess. One approach is to use empirical evidence based on previous investigations. This is dependent on the availability of a real-life case dataset and the outcomes of the investigation, i.e. the digital forensic analysis report.

3.1 Requirements for an NLP solution

The eight considerations presented in paragraph 3.1 are used to formulate requirements for an NLP solution. Stated requirements are numbered below.

Analysis-investigation cycle. The cycle implies a gradual refinement of theories, and the investigator therefore needs to see a gradual refinement in the summaries. The solution must make provision to present refined results based on inputs from the investigator.

1. Based on input from the investigator, summaries must be recalculated that must result in refined summaries.

Training time. An efficient design is required to reduce initial training time. Training time includes pre-processing. Producing initial output less accurate output may be a way to allow the investigator to start working while training continues. This can perhaps be done with inputs from the investigator even before the full training commences.

2. The model must allow for inputs to produce initial useable results to the investigator.

3. Model training must continue in the background.
4. The model must allow for streaming output during prediction.

Reducing text vs losing detail. By definition a summary of text may lead to a loss of crucial detail. The human interpreting the summary will know whether he/she needs to see more or less detail. This can be achieved by providing an input to this effect.

5. The model must allow for summary detail to be variable, and recomputeable on the fly.

Multi-theory development. Due to the statistical nature of NLP, it produces a single summary which may be biased. Summaries will inherently focus the attention of the investigator on a certain set of emails. For multi-theory development, a shift in the bias of the model is required to allow for a different summary of the emails. It is however imperative that the investigator can report on and reproduce the steps that leads to the theory development. External inputs may also be required during the multi-theory development that come from other forensic sources, or other investigative sources. Examples of input may be keywords, named entities, a time period, or even a document containing information that must be matched. It is imperative that the investigator can explain why the process and underlying model behaves in the way it does, i.e. what leads to the summaries.

6. The model must allow for a pseudo random input which leads to different summaries or views on the same set of emails.
7. The model must automatically track the steps the investigator takes for later documentation and reproducibility.
8. The model must accommodate external data which is used to prioritize the summaries produced.
9. The model must display or highlight the trigger words used to compile the summaries.

Creating timelines and finding spatial links. Emails have two natural orderings which can be leveraged. Emails are ordered in timelines by means of timestamps. This allows for a sequence of events to be constructed from emails unrelated to each other by means of time stamps. It is acknowledged that timestamps might not be fully accurate, but still allows for some time ordering. Emails have natural spatial relationships by their association between sender and receiver. Additionally, email content have additional content which can be utilized in spatial relationships, namely attachments and named entities.

10. The model must produce and link summaries by means of email trail flows.
11. The email summaries must be produced in time order.
12. Named entity recognition must be leveraged in the summaries for spatial relationships.
13. Summaries must include aspects of attachments, either by content or by attachment type.

Reporting and presentation. This is arguably the most crucial component of a solution. It is the main interface to the investigator and must make his/her life easier. Two types of reporting are required. The main reporting will be on query results, i.e. summaries. Towards this end, spatial relationships must be highlighted, timelines must be displayed, and links must be created to the original emails. A secondary reporting may be the steps taken to reach a result, i.e. historical actions taken during the queries. This is required for final reporting not only of the results, but the investigative actions in reaching the results.

14. Graphical view of email trail summaries must be produced.
15. The graphical view must highlight spatial relationships.
16. The graphical view must be produced in a time ordered manner.
17. Links to original emails must be made available for quick access.
18. Reporting on actions taken during the investigation, including search term inputs, must be produced.

Reproducibility and interpretability. This aspect was already addressed above. The models must be interpretable by highlighting important terms used by the model. The reporting function allows for the investigator to retrace his/her steps in the analysis of the data. The ability to report on the investigative process, reproduce it, and interpreting it allows for a more scientific approach in investigations whereby the investigation process can be scrutinized.

Measuring success. During model development and refinement, success factors need to be measured. A fixed testing baseline is required for this. Of crucial importance is the testing of the model in practice.

19. Freely available open datasets must be used as baseline testing.
20. Empirical testing by investigators and their observations must be documented.

The above covers all technical and investigative considerations and requirements. Finally, ethical considerations in machine learning is described next.

3.2 Ethical Considerations

The application of machine learning in any industry has ethical implications. Microsoft [11] and IEEE [12] has set off to promote the proactive addressing of ethical considerations when introducing machine learning in an industry. The application of NLP in digital forensic investigations has the side effect that decisions the investigator make is driven by a statistical learning or machine learning model. This paper identifies three ethical considerations.

Interpretability: In the United States of America, a machine learning application predicted the re-occurrence probability of offenders before granting parole or sentencing. In one such instance [10], a court challenge was lost by the state. The basis of the court challenge was that the state could not explain how the model came to that conclusion. Applying NLP in digital forensic investigations is no different. When a model is inter-

pretable, not only does it contribute to the investigator's ability to defend his/her actions, it also aids the investigator in understanding the limitation of the tool he/she is using.

Reliance on machine learning: Experienced investigators may know when to rely on their tools, and when to supplement their tools with acquired experience. Introducing a machine learning tool to inexperienced investigators, may lead to investigators in the future to only rely on the machine learning algorithms. The machine learning algorithm may inherently be biased due to its statistical nature. This must be countered by proper training programs being developed in parallel to introducing NLP techniques to investigators. It should also be considered to design proficiency tests to be taken on a regular basis to ensure that the investigators understand the nature of their tools before allowing them to use it in investigations.

Tunnel vision: One of the three main mistakes an investigator can make, is suffering from tunnel vision. It can be described as a form of bias from the investigator's perspective, i.e. starting with a biased view, and then finding evidence to support the biased view. A rigid NLP model may cause this to happen, even if an investigator tries to remain unbiased. The NLP model, due to its inherent bias, will produce summaries of texts, and therefore a biased view on the large amounts of texts. This inevitably leads to the investigator being biased without even knowing it. Reducing this bias, and being aware of it, will be up to both the NLP model designer, who has to introduce variability in generating summaries, and the investigator who has to rely on more than just one technique when extracting information from large datasets of documents.

4 General architecture to apply text summarization to emails

A general architecture for email processing is introduced in Figure 2. All elements in the architecture are conceptual elements. A process starts off with an email database containing all the seized emails with its email headers. For simplicity it is assumed that the email database is set up by existing digital forensic tools, stripping off email headers and organizing the emails into appropriate tables so that queries can be performed.

Pre-processing: Before models can be trained, the email bodies need to undergo preprocessing. During preprocessing text normalization takes place. After text normalization, named entity recognition is performed. Named entity recognition requires a special encoding to ensure that when training takes place, named entities are treated as single concepts by the NLP algorithm. Additional preprocessing steps may also include recognizing and encoding dates and times. Stop words must be removed. Part of speech tagging may also be considered as a means of extracting spatial relationships. Emails are tokenized (breaking up into words) and stored in structures whereby sentence level and paragraph level can be extracted Attachment data may optionally be encoded for the NLP model to process. The preprocessed emails are then stored in a database with links to the email header data and original email content.

Model training: The next step is model training. An initial model for topic analysis needs to be trained. A simplistic view of model training is that it produces a complex dictionary specific to the tokenized email corpus for later reference. Model training

can be an iterative process. Initial training may be once off, but to cater for variability requirements, model training may be performed more than once with random parameter selection (which is stored to satisfy reproducibility requirements). To satisfy refinement requirements, models might be retrained based on an email sub-selection of emails. Each model trained must be stored with its associated parameters (e.g. data subset selection).

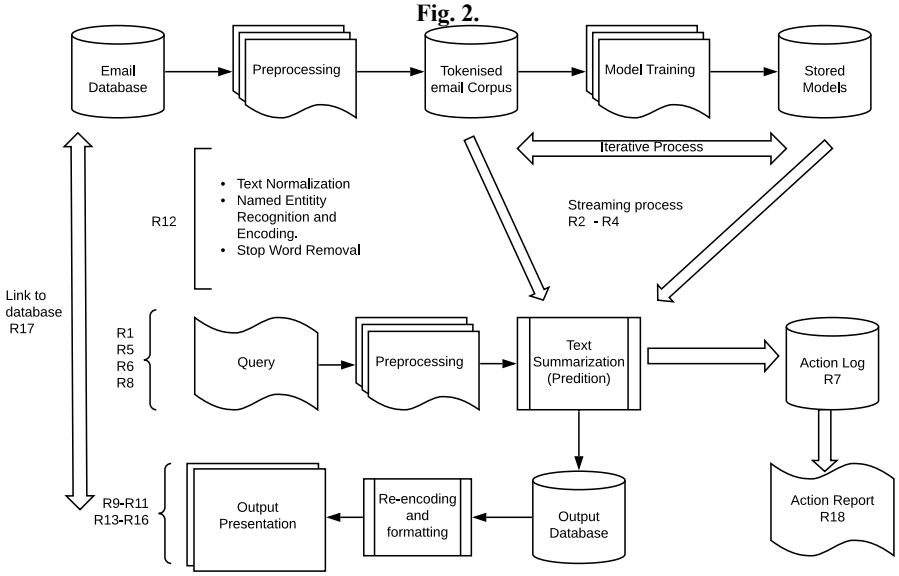
Running queries: With pre-trained models available, the digital forensic investigator can now commence performing queries. Queries described in Section 3.1 must be possible, i.e. a query can be a simple word, an entity or a document. The query itself, where phrases or documents are input as queries, will have to undergo the same preprocessing steps as the emails to ensure consistency in query matching. Where documents are added, it may be necessary to retrain models, or train new query models on the document. Note that it is allowed to run a query (prediction) on data that was not used in training the model. This allows the forensic investigator to run a query of emails related to a single document's contents for instance. Additionally, a model can be chosen to be retrained on a subset of emails, which allows the "dictionary" to be refined for more accurate queries across the entire email corpus. This introduces variability to aid in theory development and allows more flexible ways to reduce model bias. During the query phase, results may be streamed as summaries are compiled to save time.

All actions, including queries must be logged. It frees the forensic investigator from keeping notes of all actions and keeps him/her focused on the interpretation of results. It ensures an auditable log of the investigation process which can be repeated by any independent party.

Before query presentation can take place, email text must be reconstructed e.g. dates and time re-encoded to interpretable form, named entities re-encoded, word stems replaced with their original words, and sentences reconstructed. The summaries must be linked to the original email header data, ordered in time sequence and then presented. This type of feedback in itself can be a large data query and should potentially be stored for efficiency reasons. Presentation of the query results should be primarily in a timeline, and entity and topic relationships must be presented visually. Sub queries can be run on the returned results (e.g. displaying specific topics only, or filtering specific email addressed).

Each element in Figure 2 is labeled with the requirements as presented in section 3.2. This is an indication of the element where the requirements will have to be satisfied. Two requirements are not presented in the figure 2, namely requirements 19 and 20. These requirements are not satisfied by implementation aspects, but by actions. For the datasets, the Enron dataset [13] and the Hillary Clinton email dataset are two candidates to satisfy requirement 19. Requirement 20 can only be satisfied by partnering with a digital forensics company.

Fig. 2.



General architecture for email natural language processing.

5 Future Work

The architecture presented in this paper needs to be put to the test. The authors intend to develop a practical NLP model and apply it to existing openly available email datasets. Each aspect of the model can be refined, and of particular interest is the reporting and presentation function to the investigator. The model will be empirically tested by making it available to an investigator(s) working in the field of digital forensics and testing its effectiveness against actual case data. It is hoped that the evaluation by investigators will provide more insights into the technical considerations of applying NLP techniques in digital forensic investigations.

6 Related Work

In [14] the challenges of analyzing big volumes of data in forensic analysis is discussed. A two-step process is proposed for topic extraction from documents. The first step entails the text extraction, and the second step utilize clustering to group topics. Background is also provided on the actual extraction process from a technical point of view. In [14] partial documents recovered from deleted files are considered. The Enron dataset is analyzed for three email addresses. The output consists of words identified as related topics. This approach will fail on many requirements derived in this work.

Clustering and sentence extraction is used in [15] to generate summaries of documents. A measure of sentence similarity is used to define a distance function between sentences in a document. Clusters are detected and sentences are ranked to summarize

topics. This work inspired the authors of this paper in finding a better presentation than wordclouds.

A recent literature survey [16] summarizes techniques employed in text clustering for topic extraction, including optimizations that were attempted. It is a nice overview of attempts and the results of the attempts. Proper motivation is provided in [16] that current forensic tools rely on the investigator to know which search terms to use to identify the documents of interest. A good overview and comparison of all the techniques are provided.

An unpublished paper [17] focuses on emails and text summarization. The output presented to the investigator is in the form of wordclouds. The authors improved their work by taking a different approach in [18]. In Section II of [18] a requirements analysis is performed and in Section III an ontology for investigation is developed. Section IV then discusses text analysis in terms of this ontology. This promising work is perhaps the closest to our work but taking a much deeper dive into finding spatial relations.

7 Summary and Conclusions

This paper introduced investigation principles used by detectives in criminal investigations. It then introduced the basic elements of modern natural language processing. Putting the two together, this paper derived the considerations that must be made when constructing an NLP model for use in assisting investigators when analyzing large quantities of emails. Basic requirements for an NLP model was derived from the considerations and ethical considerations in applying machine learning was dealt with. A basic NLP architecture was derived that can form the basis of an NLP model applied to emails in the investigative process. Three questions were answered in this paper:

What are the main investigative and technical considerations when relying on NLP techniques in forensic investigations?

The answer to this question is covered in Section 3.1

What are the basic requirements for an NLP solution to be applied to emails in digital forensic analysis?

The answer to this question is covered by the 20 numbered requirements in Section 3.2.

What elements would a general NLP architecture consist of to process emails into useful summaries for the forensic investigator?

The answer to this question is covered in Section 4.

This paper demonstrates that the application of NLP in solving real-world problems and specifically digital forensics require a lot more than merely applying the model. Many factors need to be taken into account to find a useable model, and many refinements to the model implementation are required to make it fit for purpose.

The presented model and considerations are not complete, but merely a first attempt that can be improved on. The focus in this paper is on emails and does not cover other general texts that a forensic investigator needs to deal with.

References

1. Malan, J.: Personal Communications: Text Mining as an aid in Forensic Analysis., (2019).
2. International Standards Organisation: Information technology — Security techniques — Incident investigation principles and processes. International Standards Organisation (2015).
3. Ellerbeck, L.: Personal Communications: Aids to assist investigators in finding the source documents/topics., (2019).
4. Gehl, R., Plecas, D.: Introduction to Criminal Investigation: Processes, Practices and Thinking. Justice Institute of British Columbia (2017).
5. Ghosh, S., Gunning, D.: Natural Language Processing Fundamentals. Packt Publishing Limited.
6. Gensim Documentation and tutorials, https://radimrehurek.com/gensim/auto_examples/index.html.
7. Řehůřek, R., Sojka, P.: Software Framework for Topic Modelling with Large Corpora. In: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks. pp. 45–50. ELRA, Valletta, Malta (2010).
8. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J.: Distributed Representations of Words and Phrases and their Compositionality. In: Burges, C.J.C., Bottou, L., Ghahramani, Z., and Weinberger, K.Q. (eds.) Advances in neural information processing systems. pp. 3111–3119 (2013).
9. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient Estimation of Word Representations in Vector Space. arXiv:1301.3781 [cs]. (2013).
10. Liptak, A.: Sent to Prison by a Software Program's Secret Algorithms, https://www.nytimes.com/2017/05/01/us/politics/sent-to-prison-by-a-software-programs-secret-algorithms.html.
11. Microsoft: The Future Computed, Artificial Intelligence and its role in society. Microsoft Corporation Redmond, Washington. U.S.A. 2018 (2018).
12. IEEE: Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems. IEEE.
13. Enron Email Dataset, https://www.cs.cmu.edu/~enron/.
14. Decherchi, S., Tacconi, S., Redi, J., Leoncini, A., Sangiacomo, F., Zunino, R.: Text Clustering for Digital Forensics Analysis. In: Herrero, Á., Gastaldo, P., Zunino, R., and Corchado, E. (eds.) Computational Intelligence in Security for Information Systems. pp. 29–36. Springer Berlin Heidelberg, Berlin, Heidelberg (2009). https://doi.org/10.1007/978-3-642-04091-7_4.
15. Aliguliyev, R.M.: Clustering Techniques and Discrete Particle Swarm Optimization Algorithm for Multi-Document Summarization. Computational Intelligence. 26, 420–448 (2010). https://doi.org/10.1111/j.1467-8640.2010.00365.x.
16. Almaslukh, B.: Forensic Analysis using Text Clustering in the Age of Large Volume Data: A Review. IJACSA. 10, (2019). https://doi.org/10.14569/IJACSA.2019.0100610.
17. Spranger, M., Labudde, D.: Semantic Tools for Forensics: Approaches in Forensic Text Analysis. Draft: International Journal on Advances in Intelligent Systems. (2013).
18. Spranger, M., Labudde, D.: Towards Establishing an Expert System for Forensic Text Analysis. International Journal on Advances in Intelligent Systems. 7, 247–256 (2014).