Meta-semantic Search Engine Method Proposition for Transparent Decision Auditing

Lucas C. de Almeida[a], Francisco L. de Caldas Filho[b], Fábio L. L. de Mendonça[c]
and Rafael T. de Sousa Jr.[d]
Electrical Engineering Department, University of Brasília, Faculty of Technology, Darcy Ribeiro Campus, Brasília, Brazil

Keywords: Search Engine, Semantic Search, Meta-semantic Search, Data Enrichment, Forensics Computing.

Abstract: The use of search tools in decision-making and investigation processes has been gaining more and more space in the forensic community. The ability to index various sources of information and to be able to filter specific snippets and ideas is one of the milestones in the history of forensic and investigative computing. However, the widespread use of these methods, such as semantic search engines based on deep learning and machine learning methods can generate impractical results for complex cases. That's because the criteria these machines use to classify snippets of natural languages can be so complex that they're no longer auditable. Therefore, if a machine produces results that cannot be verified and explained, it is producing inferences that are highly questionable or even worth nullifying. In this work, we explore the advantages of applying data enrichment before the search process, and the subsequent use of keyword search tools to present an indexing framework with more transparent criteria and more practical results for the defense of ideas based on the findings from the use of the tools.

1 INTRODUCTION

According to the Cybersecurity and Infrastructure Security Agency of the United States (CISA), computer forensics is defined as ”the process of using scientific knowledge for collecting, analyzing, and presenting evidence to the courts” (CISA, 2008). In modern processes, this stage of the investigation process becomes even more important, since most of the communication and information recording processes may take place digitally, whether through the use of cell phones, computers, instant messaging applications, electronic mails, audio messages, among many others. One can see some history about it in (Britannica, 2022). In cases of bank fraud, identification of fraud gangs, scams against property, businesses carried out digitally in an unreliable way, to name just a few scenarios, the role of data collection, organization, cleaning, filtering and construction of theses is extremely necessary for there to be a case to be judged or even for the decision to open an investigation. And this role only tends to grow and become more popular, given that, today, business institutions and commercial relationships tend to rely and focus much more on digital than personal. One can verify an example of it in (Nielsen, 2012). That trend has led to an unprecedented expansion of the use of technology as a form and means of expression, sustenance and association between individuals, (Archibugi and Iammarino, 2002) has a complete work on the subject.

Therefore, the more popular and widespread the use of computer forensics in investigation processes, the larger and more complex the data repositories to be analyzed can become. That is, if a process requires the analysis of data from different sources, in different formats, and in massive amounts, as could occur during the monitoring and identification of a gang of credit card fraud, for example, more relevant and common the concepts of Big Data, Data Warehouse and Data Lake may become for investigators. This occurs naturally, similar to the fact that the general process of computer forensics, which is well explained (University, 2017), is very similar to the OSEMN framework. The work in (Almeida et al., 2021) presents a good project based on that framework. Ideally, investigators should be able to transform a Data Lake, defined as unstructured data from a variety of sources (Oracle, 2022), into a Data Warehouse, de-

[a] https://orcid.org/0000-0002-2519-1574
[b] https://orcid.org/0000-0001-5419-2712
[c] https://orcid.org/0000-0001-7100-7304
[d] https://orcid.org/0000-0003-1101-3029

C. de Almeida, L., Filho, F., L. de Mendonça, F. and Sousa Jr., R.
Meta-semantic Search Engine Method Proposition for Transparent Decision Auditing.
DOI: 10.5220/0011356700003280
In *Proceedings of the 19th International Conference on Smart Business Technologies (ICSBT 2022)*, pages 80-89
ISBN: 978-989-758-587-6; ISSN: 2184-772X

fined as structured and well-behaved data for generating insights (MIT, 2022), using techniques for Big Data processing, that can be described as data that is so large or varied that cannot be processed using conventional methods (SAS, 2022). However, one of the biggest challenges relies on building a coherent database with clear and concise conclusions that are unbiased and auditable. The conclusion is straightforward, therefore, that the use of completely autonomous processes and with decision criteria originated from iterative numerical methods can be questioned both because of the possible biases it presents and because of the difficulty in verifying step-by-step how each decision has been made and how it compare to similar cases.

Still on the problems and limitations of the use of completely autonomous decision machines, there are scenarios in which their use can end up inferring very expressive and harmful errors for the investigation process. This is due to the fact that machines are still not completely capable of understanding sarcasm and intentionally malicious substitutions. In (Katyayan and Joshi, 2019), one can have a general overview on the subject. A clear example of this concept is seen in the monitoring of conversations of criminal groups. It is quite common to try to deceive systems and investigators using Figures of speech and exchanging names and verbs that would clearly indicate crimes for expressions that leave several possible meanings open. That is, criminals tend to create codes and communication patterns that resemble processes that are completely unrelated to the criminal activity they are associating with. In this sense, the harm of using semantic search based on machine learning or deep learning in a context of large Data Lakes can produce results with little or no value, and let conversations and entire evidence go unnoticed that, with more careful search, could be valuable objects in the investigation. Therefore, how to build a research thesis based only on the result of correlations found using methods and algorithms of extreme complexity and based on probabilistic heuristics with its values and weights hidden (or still impossible to understand) inside the machine?

A last relevant point to be mentioned is that the uncontrolled use of algorithms and decision machines based on deep learning and machine learning methods can make it very difficult or almost impossible to measure the similarity between the results of different investigative processes. This is because usually there is no friendly understanding of the weights and formulas generated by such a machine, as seen in (MIT, 2018). Just as courts tend to judge based on past decisions, following the basic concept of jurisprudence (Cornell, 2022a), and these decisions can be audited if it turns out that they do not follow the normality of decisions in similar cases, investigators should also be able to study and audit the result of their work in order to produce evidence (i.e. data and theses) with a similar profile for cases of the same nature or with similar sources of information, including being able to verify and measure the bias of the filtered data independently of the decisions of courts.

A possible (and widely used) alternative to semantic search is keyword search. In this type of search, the researcher freely chooses words and excerpts from possible conversations that he/she finds relevant and the system searches, enumerates, presents and counts the times in which exact reproductions of the terms (or keys) occurred in the data sources. One can see a famous example of a keyword search engine in (Luo et al., 2008). The problem with this approach is that it becomes extremely difficult to predict and/or enumerate all possible terms and expressions relevant to a case, specially if these choices are the responsibility of a human being, who will have limited and biased vocabulary. The work in (Lynn K, 2022) shows some interesting findings about that concept. However, this method has advantages, the main one being based on the fact that the results produced are auditable, comparable with different scenarios and data sources, and do not present any bias beyond that already existing in the keys used for the search.

An interesting approach, therefore, would be the association of the two methods. That is, if there were a way to treat data with complex and probabilistic methods and heuristics, but at the end of the treatment, search using keywords and specific expressions, it would be possible to build a machine that helps researchers to produce evidence with measurable bias, auditable results and comparable with other similar cases, and still with simple heuristics easy to be used in theses in front of a court.

The purpose of this work, therefore, is to propose a generic meta-semantic search method that produces auditable and comparable results with other application cases. It happens that there are different ways to build a machine that performs searches with semantic content, and such alternatives operate without necessarily involving algorithms and iterative numerical processes in the final decision. Therefore, unlike a machine that performs semantic searches with complex heuristics, it is possible to search (using Big Data techniques) for keywords and specific expressions in databases enriched with the use of algorithms and data enrichment machines with an emphasis on semantics. In Figure 1, one can understand the basic difference between the two approaches. The relevance of such framework can be better understood in the Methodology section.

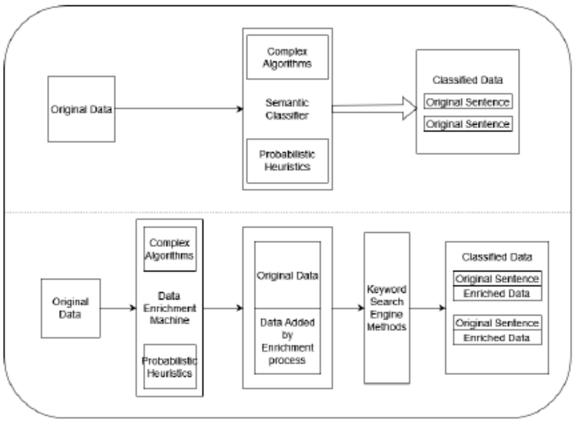


Figure 1: Two approaches to semantic search. The one on top is the usual method relying in semantic classifiers. The one on the bottom shows a search procedure that produces meta-semantic search results.

As a validation of the proposed framework, it was possible to set up an experiment in which a process of semantic enrichment was applied to a list of words for the subsequent search using keywords. It produced results that were compared with those obtained by the direct application of a semantic classifier, concluding on the feasibility of the framework, which presents itself as an interesting alternative for the construction of such forensic systems, as will be presented throughout this document.

This paper consists of five sections, including this introduction. Section 2 deals with related work, especially the necessary technical concepts. Section 3 specifies the proposed solution, uniting method- ologies widely used in the industry and that integrate seamlessly into the context of users. Section 4 serves to discuss and demonstrate the results obtained in implementing the project with real data. Findings and future work can be found in Section 5.

2 RELATED WORK AND PRIOR ART

The usage of digital platforms such as social media and instant messaging applications to support criminal activites has been gaining a lot of attention in the present of this work, as can be seen in (Affairs, 2021) and (Forbes, 2021). The fact that the courts and governments want Internet companies to cooperate an provide data is a sign of the growing need of more information to support investigations and processos judiciais.

In (Mukhopadhyay et al., 2013), there is a proposal for a meta-semantic search tool for Internet content. It is interesting to note that the concepts that involve the construction of new propositions for searching and indexing information with semantic basis were well constructed, however, it would be an interesting advance to be able to audit, at least partially, the internal mechanisms of data classification, such as is proposed in the current work.

In (Soltani et al., 2021), the importance and role of the use of search engines and data indexing in the process of computer forensics is well explored, representing an excellent review of the main methods used and problems encountered in the course of an investigation. However, unlike the current work, the focus is only on the technical concepts that involve the investigation, and an evolution is still necessary in the questioning of the validity and transparency of the results of the application of such methods in front of a court. Yet, there is little concern about the need to produce neutral evidence with measurable bias and comparable to other cases.

Also, in (Gutiérrez et al., 2016) it is possible to understand and observe examples of data enrichment and how this process can be valuable for different branches of Industry. The same concept is applied in current work in an attempt to make the investigation of large masses of data more efficient and transparent, in addition to facilitating the creation of theses and the inclusion of similar cases as support. In other words, the use of data enrichment processes in a non-hidden way allows legal thesis to be assembled and lines of reasoning to be created. As an example, even if two groups of sentences of two cases are very different (perhaps in an attempt to obfuscate an illicit activity), in a past case the result of enriching a group of sentences resulted in the aggregation of several words and terms quite similar to those verified in a present case, and if in the past case the final result of the investigations proved that an individual was part of a gang, it would be plausible to ask for an investigation to be opened (or to hold a business transaction, or even to start a police operation) in a case that is visibly similar to the past one.

Finally, in (Di Nunzio, 2004) the use, classification and plotting of words and phrases as vectors in two dimensional planes is demonstrated in a simple and efficient way to help in document classifications. These concepts form the basis of motivation for the proposition and validation of the meta-semantic search framework described in the current work.

In the following section, the technical details of an experiment created for validating the proposed framework will be described.

3 METHODOLOGY

It is interesting to understand, First, if the proposed framework can, in an analytical way, have better results or at least equivalent to those related to the use of a generic classifier. In addition, one should explore what would be expected from a scenario involving criminal activities, which are the focus of this model.

Based on a conversation between common users on any subject, for example a sale at a convenience store, it is in the interest of both communicators to be clear about their intentions and opinions so that their business relationship has the desired end efficiently. Bearing this premise in mind and using the approach of describing words and sentences as vectors, as exemplified in (CHURCH, 2017), the application of a generic binary classifier with its decision boundary can be seen in Figure 2. The green dots in the Figure are sentences, words or expressions plotted (or projected) on a two-dimensional plane using whatever criteria are relevant to the generic classifier. The red region would be the one in which the classifier will have the maximum probability of considering any points as having the same sense, and the yellow region is where the so-called decision boundary of the classifier would be (defined as the boundary that separates two classes under the point of view of the machine). In this region, the probability of a point being misclassified is quite high and varies discontinuously, that is, points may be misclassified more often compared to any other part of the Figure. In this case, where the data is generally well behaved (since the actors in the conversation act intentionally efficiently), a well-built and optimized classifier could have a very satisfactory success rate.

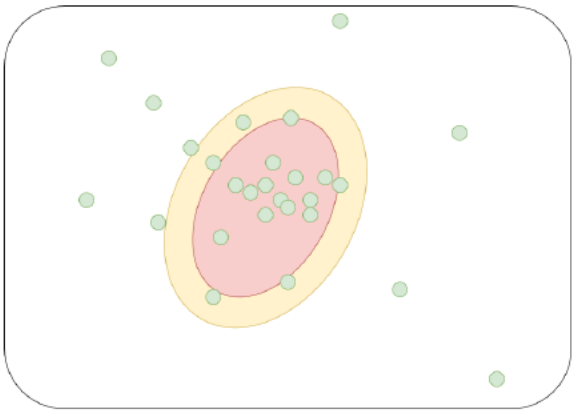


Figure 2: Common users conversation with words and expressions projected in a two dimensional generic plane and with a generic binary classifier applied to that plane.

However, it should be remembered, as mentioned in the Introduction of this work, that the communication of criminals tends to be done in an obfuscated way, that is, they tend to change names, expressions, phrases and meanings, all with the objective of deceiving systems and reducing the chance of commitment before courts and investigations. The effect, therefore, of the criminals' efforts acts in an equivalent way to the vectors in Figure 3, bringing a lot of bias to the data and resulting in low classification quality even for optimized machines, as can be seen in Figure 4, where the purple dots are the result of the previous green ones after the action of the obfuscating vectors.

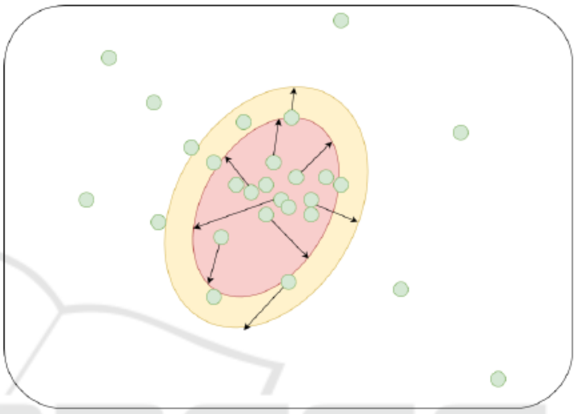


Figure 3: Conversation of criminals with a tendency to overshadow their illicit acts. The green dots with arrows are the expressions and words being obfuscated, which eventually will fall near (or outside) the decision boundary of the classifier, leading to low quality results.

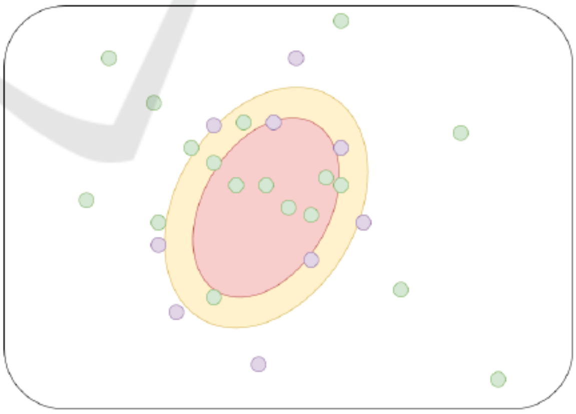


Figure 4: The result of the criminals' attempt to obfuscate their conversations generates points in the plane that fall out of or tend to stay on the edge of the decision boundary of a machine classification algorithm, a region in which the theoretical explanation and verification of the parameters used in the decision tend to be unclear and behave discontinuously.

Although more complex classification methods such as those based on deep learning, which are present in an overview discussed in (Minaee et al., 2021), can produce discontinuous classification regions, the same phenomenon described will be observed: speakers will be able to confuse classification machines and data may be discarded without a more careful look (which may be unfeasible for Big Data contexts).

In a different way, in the application of the framework proposed in this work, in which the data is First enriched for the subsequent search for keywords, the effect of obfuscation created by criminals can be mitigated. This is because, although a group of individuals uses slang and expressions with hidden meanings, there are hidden patterns in the sentences and questions that are related to the structure of the language itself. The discussion in (Picard et al., 2013) comments on the subject. These patterns will hardly be overshadowed, since the speakers of the conversation themselves would no longer be able to communicate efficiently.

A machine that, therefore, uses a previously validated and classified database, and inflates the new data using the existing ones as a parameter, if it performs this task recursively (that is, using the results of each iteration to increase the scope of the enrichment of the next iteration), would cultivate a tendency for previously expected words and expressions that did not exist in the original data to exist in the now enriched dataset. This way, an investigator would be able to perform direct searches using keywords and expressions. This trend becomes even more evident and logical if the data used for enrichment has been previously classified and validated. In practical terms, it would be the equivalent of an investigation department that has already stored several cataloged conversations of criminals in a certain field. During the investigation of individuals suspected of working in the same field, investigators can ask a machine to enrich this data with pre-existing data from successful (or unsuccessful) investigations. After enrichment, they can search for specific terms and verify if the keys were found in the original texts or in the inflated ones, and with the filtered data, they can evaluate if these are biased by comparing the inflated data with those of other cases or between those of the same scenario. This would allow them to measure how strong the evidence is or not, whether the inflated standards are adherent to the investigation, whether certain results should be discarded as if they are outside the focus of the work, among several other benefits.

Figure 5 shows the minimum expected result of a data enrichment process when the words and expressions are projected in a plan. The system generates an increase in the ”footprint” of a given point by agglutinating other similar points in the vicinity of the original point. If this process is repeated recursively, regions with a lot of enrichment and others with little enrichment should be observed, that is, regions with many blue dots, and others with few or almost none. After analyzing the results of keyword searches, an investigator could conclude, for example, that a given dataset is not sufficiently adherent to a branch of criminal activity (because it presents little inflated data), and that another dataset may produce much more results using keywords and that the recursive inflationary process ends up producing much more inflated data, meaning a more clear match with that kind of criminal activity.

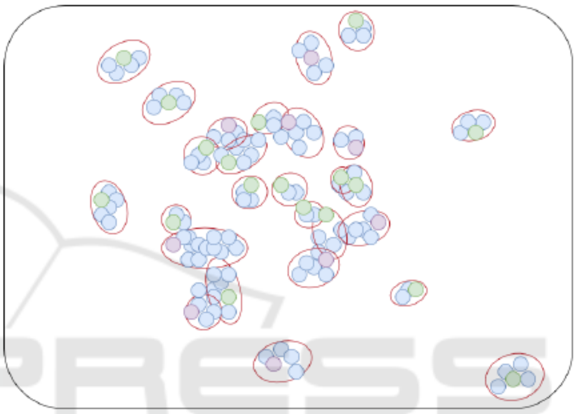


Figure 5: The green and purple dots are the original vectors related to the criminal conversations (purple are the ones moved by obfuscating vectors). The blue dots represent data that was added by a generic semantic classifier process. The circled area marks the clusters formed by the data agglutination.

In the Figure 6 it is possible, finally, to understand the purpose of the proposed framework. While a generic classifier looks for regions and tries to coalesce the original data using non-transparent methods, the proposed method inflates existing data so that a user can rely on simple terms and phrases and still have a high chance of finding these keys in inflated data clusters. A well-placed allusion resembles an attempt to increase the targets on a target shooting stand so that the challenger has a better chance against the stand’s owners. In a normal situation, the stand’s owners (the criminals) have the advantage (the data is obfuscated and disconnected), but with the help of the method described in this work, the client that challenges the stand (the investigators), without increasing the darts, can find it easier to hit the targets, which are now many times larger and, in some cases, even merged with each other. One can now compare the expected results in Figure 6 and Figure 4.

The objective of this work is, therefore, to sim-

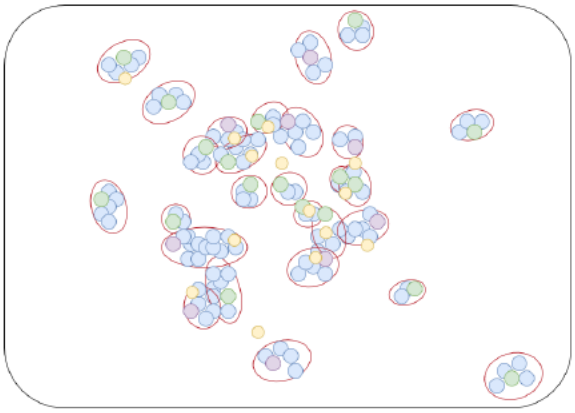


Figure 6: Green, purple and blue dots forming clusters, and yellow dots representing the keyword searches the investigator may perform.

ulate the application of an alternative framework for searches with semantic effect using keywords in large data repositories, being especially useful for forensic investigations. After the simulation, it will be possible to compare the result of applying the model in relation to the use of a semantic classification machine with complex methods operating internally. The framework to be tested will be based on the one shown at the bottom of Figure 1, that is, an initial natural language database will be enriched and later indexed by a simple keyword search engine, from which searches will be made for specific terms using both the original data and the inflated data as parameters.

To provide a final and clear explanation on the name of the document, the reason why the proposed method would be a meta-semantic search process is related to the fact that the results are deterministic based on the enriched data, but the enrichment process is based on semantic probabilistic heuristics, so the final findings will be indirectly related to such machine semantic classifications.

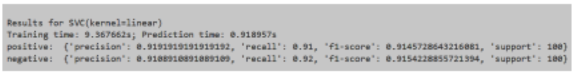
The experiment proposed to validate the framework consisted of the sentiment classification test of movie reviews using a famous dataset available at (Cornell, 2022b). The following steps were performed:

1. First, a sentiment analysis was performed using a support vector machine (SVM) with a linear kernel function, which is explained in (MIT, 1997), to obtain the rating of some movie reviews, whether the evaluation would have a positive or negative opinion about a movie. The vectorization of the reviews was done using the Term Frequency Inverse Document Frequency algorithm, commonly called TF-IDF, explained in (Havrlant and Kreinovich, 2017);
2. With the results of the classification using the support vector machine, a dataframe was created in which each review would be accompanied by the sentiment attributed by the SVM in the previous step;
3. Some reviews were chosen randomly by the system to form a base that would correspond to the ’original data’, the main target of a possible investigation into the sentiment of these using the proposed framework in this document;
4. All reviews classified in the same way (or same class, positive or negative) by the support vector machine on the First step were inserted as enrichment data for the selected reviews, those chosen to form the ”original data”;
5. Searches for specific terms were performed, and if the program found an exact occurrence of the term or expression used, either in the selected ”original” reviews or in the inflated data, it would return the results. So our goal would be to search for specific words and find same or close results to the ones found by the SVM in the First step, but now we would be able to clearly verify and judge why we got those results analyzing the inflated data.

Next, the tests and results of the implementation of the proposed experiment will be presented.

4 RESULTS

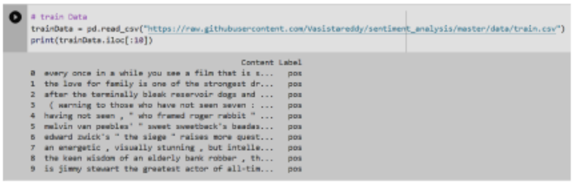
The First part of the experiment consisted of training a support vector machine to classify the movie reviews in ”positive” or ”negative” sentiment about the movies they were referring to. Since SVM is a supervised machine learning algorithm, it was possible to measure the precision of the classifier, as can be seen in Figure 7. Also, the First ten reviews can be seen on Figure 8.



```
Results for SVC(kernel=linear)
Training time: 9.367662s; Prediction time: 0.918957s
positive:  {'precision': 0.9191919191919192, 'recall': 0.91, 'f1-score': 0.9145728643216081, 'support': 100}
negative:  {'precision': 0.9108910891089109, 'recall': 0.92, 'f1-score': 0.9154228855721394, 'support': 100}
```

Figure 7: Classification report of the trained SVM model for natural language processing based on the movies review dataset. The machine reached almost 92% precision for positive reviews and close to 91% precision for negative reviews.

After the fourth step of the experiment, it was found that using all classified reviews during training to inflate the selected data ended up exaggerating the size of the clusters, causing overfitting in the inflated data, concept explained in (IBM, 2022). This resulted in searches for keywords and expressions without any value, as the chance of exact occurrences of simple

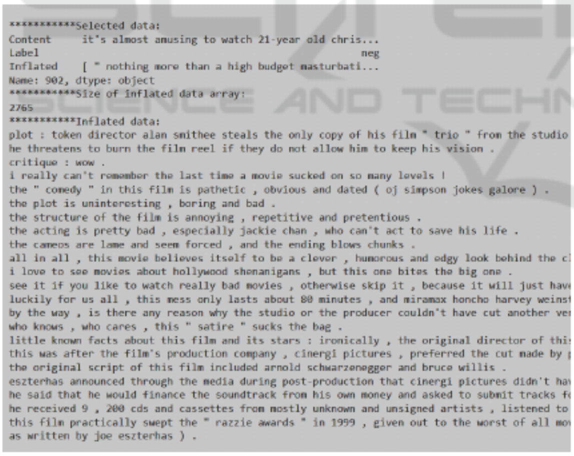


```
# train Data
trainData = pd.read_csv("https://raw.githubusercontent.com/Vasistareddy/sentiment_analysis/master/data/train.csv")
print(trainData.iloc[:10])

                                             Content Label
0  every once in a while you see a film that is s...   pos
1  the love for family is one of the strongest dr...   pos
2  after the terminally bleak reservoir dogs and ...   pos
3   ( warning to those who have not seen seven : ...   pos
4  having not seen , " who framed roger rabbit " ...   pos
5  melvin van peebles' " sweet sweetback's baadas...   pos
6  edward zwick's " the siege " raises more quest...   pos
7  an energetic , visually stunning , but intelle...   pos
8  the keen wisdom of an elderly bank robber , th...   pos
9  is jimmy stewart the greatest actor of all-tim...   pos
```

Figure 8: First ten rows of the training part of the movies review dataset, which had a copy downloaded from (Reddy, 2018).

terms and expressions became very high. That is, the ”footprint” of the selected data became so large after enrichment that the two classes overlapped. In an attempt to reduce overfitting, the number of enrichment iterations was reduced to 100. That is, instead of inflating the data with all the reviews that had been classified in the same way, the agglutination of data was limited to only 100 reviews considered similar for each class. From this point on, more significant and more frequent results began to be verified, and after a few attempts, satisfactory findings were obtained, as can be seen in Figures 10 to 14. In figure 9 it can be seen one example of part of an inflated row of the select data. Note that the printed results also show data from the inflated portion when the keyword or expression was found there, and also the index of the selected reviews from their original dataset.

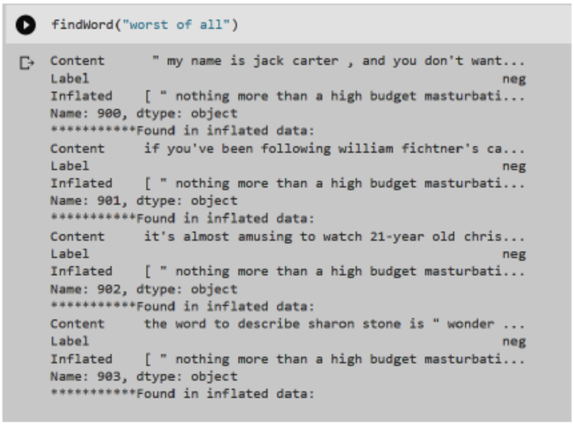


```
***********Selected data:
Content     it's almost amusing to watch 21-year old chris...
Label                                                    neg
Inflated    [ " nothing more than a high budget masturbati...
Name: 902, dtype: object
***********Size of inflated data array:
2765
***********Inflated data:
plot : token director alan smithee steals the only copy of his film " trio " from the studio
he threatens to burn the film reel if they do not allow him to keep his vision .
critique : wow .
i really can't remember the last time a movie sucked on so many levels !
the " comedy " in this film is pathetic , obvious and dated ( oj simpson jokes galore ) .
the plot is uninteresting , boring and bad .
the structure of the film is annoying , repetitive and pretentious .
the acting is pretty bad , especially jackie chan , who can't act to save his life .
the cameos are lame and seem forced , and the ending blows chunks .
all in all , this movie believes itself to be a clever , humorous and edgy look behind the cl
i love to see movies about hollywood shenanigans , but this one bites the big one .
see it if you like to watch really bad movies , otherwise skip it , because it will just have
luckily for us all , this mess only lasts about 80 minutes , and miramax honcho harvey weinst
by the way , is there any reason why the studio or the producer couldn't have cut another ver
who knows , who cares , this " satire " sucks the bag .
little known facts about this film and its stars : ironically , the original director of this
this was after the film's production company , cinergi pictures , preferred the cut made by
the original script of this film included arnold schwarzenegger and bruce willis .
eszterhas announced through the media during post-production that cinergi pictures didn't hav
he said that he would finance the soundtrack from his own money and asked to submit tracks fo
he received 9 , 200 cds and cassettes from mostly unknown and unsigned artists , listened to
this film practically swept the " razzie awards " in 1999 , given out to the worst of all mov
as written by joe eszterhas ) .
```

Figure 9: Part of a inflated row of the selected data dataframe built for the fourth step of the experiment.

In Figure 14, it is possible to see that the search for extremely simple and direct terms may tend not to produce results with relevant quality. In the example, searching for the simple term ”good” returned results of several positive and negative reviews.

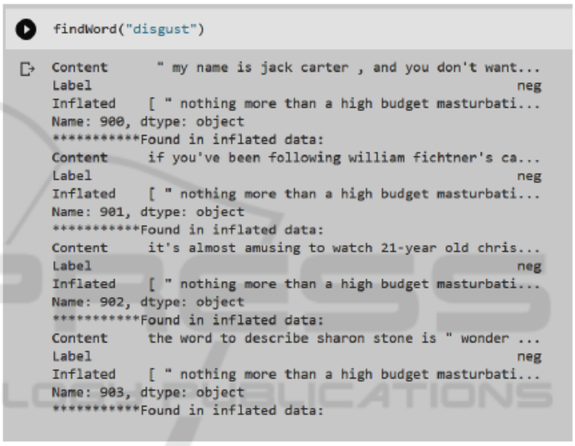
A similar behavior can also be observed when a search for simple words and expressions of extreme positive sentiment are found in the enrichment (or in the original data) of reviews with sentiment classified as negative. This happens on two possible occasions:



```
findWord("worst of all")

Content      " my name is jack carter , and you don't want...
Label                                                     neg
Inflated     [ " nothing more than a high budget masturbati...
Name: 900, dtype: object
***********Found in inflated data:
Content     if you've been following william fichtner's ca...
Label                                                     neg
Inflated     [ " nothing more than a high budget masturbati...
Name: 901, dtype: object
***********Found in inflated data:
Content     it's almost amusing to watch 21-year old chris...
Label                                                     neg
Inflated     [ " nothing more than a high budget masturbati...
Name: 902, dtype: object
***********Found in inflated data:
Content     the word to describe sharon stone is " wonder ...
Label                                                     neg
Inflated     [ " nothing more than a high budget masturbati...
Name: 903, dtype: object
***********Found in inflated data:
```

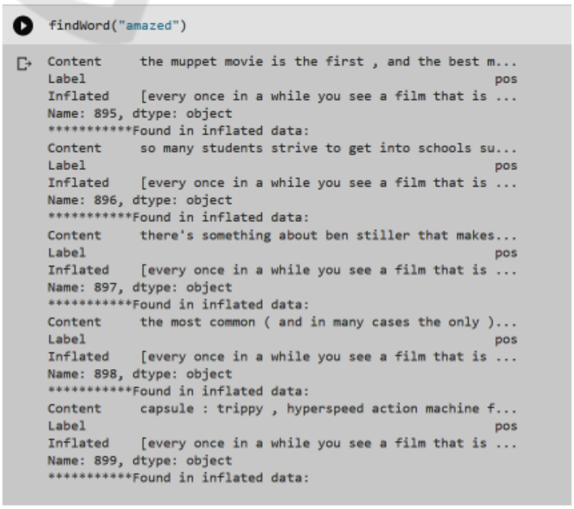
Figure 10: Result of the search for exact matches for the expression ”worst of all”. Only negative reviews were found (aligned with the ”Label” line).



```
findWord("disgust")

Content      " my name is jack carter , and you don't want...
Label                                                     neg
Inflated     [ " nothing more than a high budget masturbati...
Name: 900, dtype: object
***********Found in inflated data:
Content     if you've been following william fichtner's ca...
Label                                                     neg
Inflated     [ " nothing more than a high budget masturbati...
Name: 901, dtype: object
***********Found in inflated data:
Content     it's almost amusing to watch 21-year old chris...
Label                                                     neg
Inflated     [ " nothing more than a high budget masturbati...
Name: 902, dtype: object
***********Found in inflated data:
Content     the word to describe sharon stone is " wonder ...
Label                                                     neg
Inflated     [ " nothing more than a high budget masturbati...
Name: 903, dtype: object
***********Found in inflated data:
```

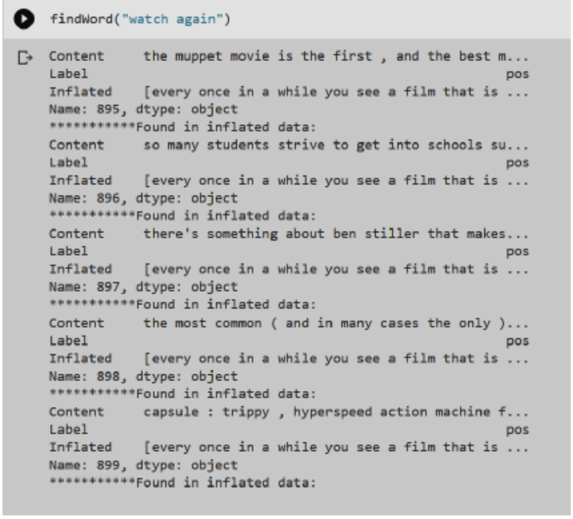
Figure 11: Result of the search for exact matches for the expression ”disgust”. Only negative reviews were found (aligned with the ”Label” line).



```
findWord("amazed")

Content     the muppet movie is the first , and the best m...
Label                                                     pos
Inflated    [every once in a while you see a film that is ...
Name: 895, dtype: object
***********Found in inflated data:
Content     so many students strive to get into schools su...
Label                                                     pos
Inflated    [every once in a while you see a film that is ...
Name: 896, dtype: object
***********Found in inflated data:
Content     there's something about ben stiller that makes...
Label                                                     pos
Inflated    [every once in a while you see a film that is ...
Name: 897, dtype: object
***********Found in inflated data:
Content     the most common ( and in many cases the only )...
Label                                                     pos
Inflated    [every once in a while you see a film that is ...
Name: 898, dtype: object
***********Found in inflated data:
Content     capsule : trippy , hyperspeed action machine f...
Label                                                     pos
Inflated    [every once in a while you see a film that is ...
Name: 899, dtype: object
***********Found in inflated data:
```

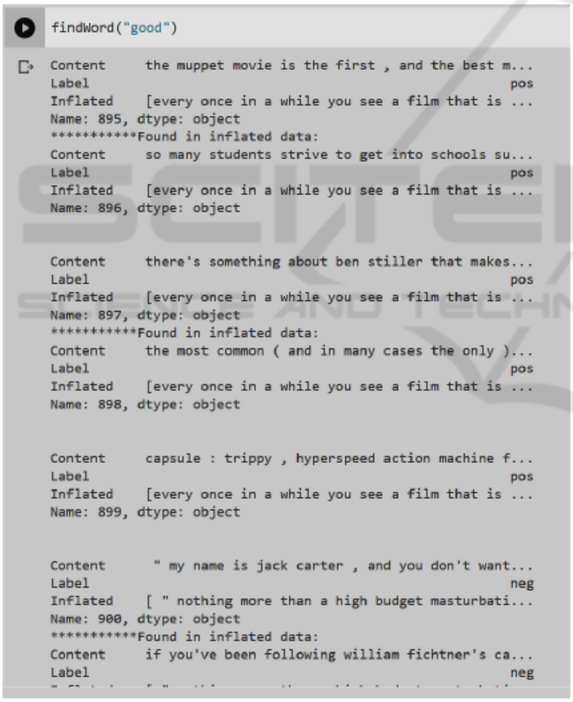
Figure 12: Result of the search for exact matches for the expression ”amazed”. Only positive reviews were found (aligned with the ”Label” line).



```
findWord("watch again")

Content     the muppet movie is the first , and the best m...
Label                                                    pos
Inflated    [every once in a while you see a film that is ...
Name: 895, dtype: object
***********Found in inflated data:
Content     so many students strive to get into schools su...
Label                                                    pos
Inflated    [every once in a while you see a film that is ...
Name: 896, dtype: object
***********Found in inflated data:
Content     there's something about ben stiller that makes...
Label                                                    pos
Inflated    [every once in a while you see a film that is ...
Name: 897, dtype: object
***********Found in inflated data:
Content     the most common ( and in many cases the only )...
Label                                                    pos
Inflated    [every once in a while you see a film that is ...
Name: 898, dtype: object
***********Found in inflated data:
Content     capsule : trippy , hyperspeed action machine f...
Label                                                    pos
Inflated    [every once in a while you see a film that is ...
Name: 899, dtype: object
***********Found in inflated data:
```

Figure 13: Result of the search for exact matches for the expression ”watch again”. Only positive reviews were found (aligned with the ”Label” line).



```
findWord("good")

Content     the muppet movie is the first , and the best m...
Label                                                    pos
Inflated    [every once in a while you see a film that is ...
Name: 895, dtype: object
***********Found in inflated data:
Content     so many students strive to get into schools su...
Label                                                    pos
Inflated    [every once in a while you see a film that is ...
Name: 896, dtype: object


Content     there's something about ben stiller that makes...
Label                                                    pos
Inflated    [every once in a while you see a film that is ...
Name: 897, dtype: object
***********Found in inflated data:
Content     the most common ( and in many cases the only )...
Label                                                    pos
Inflated    [every once in a while you see a film that is ...
Name: 898, dtype: object


Content     capsule : trippy , hyperspeed action machine f...
Label                                                    pos
Inflated    [every once in a while you see a film that is ...
Name: 899, dtype: object


Content      " my name is jack carter , and you don't want...
Label                                                    neg
Inflated    [ " nothing more than a high budget masturbati...
Name: 900, dtype: object
***********Found in inflated data:
Content     if you've been following william fichtner's ca...
Label                                                    neg
```

Figure 14: Result of the search for exact matches for the expression ”good”. Both positive and negative reviews were found, and in relevant amounts for each side.

the semantic classifier may have been wrong or the reviewer may have praised some aspect of the film or actor, despite the general evaluation being negative. These cases are very similar to what would be expected from a criminal conversation with obfuscation, making the benefit of the proposed framework clear: the searcher is able to study and audit results to easily evaluate false positives and/or negatives.

With the presentation of the results of the proposed framework for the investigation around sentiment related to a movies reviews dataset, the next section gives a brief overview of the future possibilities of this application and concludes the work.

5 CONCLUSION AND FUTURE WORKS

The adoption of the proposed framework may still remain as a challenge. Although the use and general application of it for existing algorithms and methods may be simple, and the results, when well behaved, may be really significant, it is needed a previous knowledge from the searcher about the dataset and the wanted results, so the words and expressions searched fit well with the inflated data. This may not be a problem for a forensics investigator, but for the general user, it would require much effort for benefits that are specific for analysis and investigation (or thesis derivations) purposes.

Despite the difficulty of implementation in more complex scenarios, it was possible to observe that the results obtained were easy and quick to evaluate. Furthermore, the findings were completely auditable, as it was possible to track the exact location of the search keys and with some additional effort, it would even be possible to track which review was used to generate that line of the inflated data array which contributed for a given keyword search.

It is important to note that there was a relationship between the amount of data added during the enrichment process and the accuracy of the results. Therefore, even though it was a simple adjustment, a small optimization step was necessary in order to tune the enrichment and improve the class separation to avoid overfitting. In Big Data contexts, this procedure could require much more effort and study on the data, or the development of automated optimization techniques.

It should also be remembered that if the data source used for enrichment is the same in different cases, it is quite quick and easy to compare the results produced by the investigation process in both scenarios, allowing the researcher to be more faithful and critical in relation to different investigations outcomes.

The conclusions presented direct the project, finally, to a maturation in the automation point of view. Therefore, although it is interesting to facilitate the investigator's work and make the evidence found more user-friendly and auditable, it would be interesting to add another machine in the process that was capable of searching for keywords and expressions with the same semantic value as those chosen by the investigator. That is, instead of searching only for the exact occurrence of the term that the researcher chooses, the system could search for exact occurrences of synonyms and/or expressions similar to those that the researcher chose, and rank the results from such inferences with lower scores in relation to those that present a direct occurrence of the typed term. It would also be extremely important to compare the proposed framework and the application of machine learning and deep learning algorithms using a dataset containing conversations and data directly related to crimes, so that the obfuscation attempts were more evident.

ACKNOWLEDGEMENTS

This work is supported in part by CNPq - Brazilian National Research Council (Grants 312180/2019-5 PQ-2 and 465741/2014-2 INCT on Cybersecurity), in part by the Administrative Council for Economic Defense (Grant CADE 08700.000047/2019-14), in part by the General Attorney of the Union (Grant AGU 697.935/2019), in part by the National Auditing Department of the Brazilian Health System SUS (Grant DENASUS 4/2021), in part by the General Attorney's Office for the National Treasure (Grant PGFN 23106.148934/2019-67), and in part by the University of Brasilia (Grant 7129 FUB/EMENDA/DPI/COPEI/AMORIS).

REFERENCES

Affairs, S. (2021). Security affairs - telegram is becoming the paradise of cyber criminals. https://securityaffairs.co/wordpress/122609/cyber-crime/telegram-cybercrime.html.

Almeida, L. C. d., Filho, F. L. d. C., Marques, N. A., Prado, D. S. d., Mendonça, F. L. L. d., and Sousa Jr., R. T. d. (2021). Design and evaluation of a data collector and analyzer to monitor the covid-19 and other epidemic outbreaks. In Rocha, Á., Ferrás, C., López-López, P. C., and Guarda, T., editors, *Information Technology and Systems*, pages 23–35, Cham. Springer International Publishing.

Archibugi, D. and Iammarino, S. (2002). The globalization of technological innovation: Definition and evidence. *Review of International Political Economy*, 9(1):98–122.

Britannica (2022). Britannica - acquisition and recording of information in digital form. https://www.britannica.com/technology/information-processing/Acquisition-and-recording-of-information-in-digital-form.

CHURCH, K. W. (2017). Word2vec. *Natural Language Engineering*, 23(1):155–162.

CISA (2008). Cybersecurity and infrastructure security agency of the united states (cisa) - computer forensics. https://www.cisa.gov/uscert/sites/default/files/publications/forensics.pdf.

Cornell (2022a). Cornell - jurisprudence. https://www.law.cornell.edu/wex/jurisprudence.

Cornell (2022b). Cornell - movie review data. http://www.cs.cornell.edu/people/pabo/movie-review-data/.

Di Nunzio, G. M. (2004). A bidimensional view of documents for text categorisation. In McDonald, S. and Tait, J., editors, *Advances in Information Retrieval*, pages 112–126, Berlin, Heidelberg. Springer Berlin Heidelberg.

Forbes (2021). Forbes - how the fbi unmasked a whatsapp and whisper user in a pedophile sting. https://www.forbes.com/sites/thomasbrewster/2021/04/12/how-the-fbi-unmasked-a-whatsapp-and-whisper-user-in-a-pedophile-sting/?sh=3ac6e12641b5.

Guti érrez, Y., Vázquez, S., and Montoyo, A. (2016). A semantic framework for textual data enrichment. *Expert Systems with Applications*, 57:248–269.

Havrlant, L. and Kreinovich, V. (2017). A simple probabilistic explanation of term frequency-inverse document frequency (tf-idf) heuristic (and variations motivated by this explanation). *International Journal of General Systems*, 46(1):27–36.

IBM (2022). Ibm - what is overfitting? https://www.ibm.com/cloud/learn/overfitting.

Katyayan, P. and Joshi, N. (2019). *Sarcasm Detection Approaches for English Language*, pages 167–183. Springer International Publishing, Cham.

Luo, Y., Wang, W., and Lin, X. (2008). Spark: A keyword search engine on relational databases. In *2008 IEEE 24th International Conference on Data Engineering*, pages 1552–1555.

Lynn K, P. (2022). Lynn k, perry - the shape of the vocabulary predicts the shape of the bias. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3222225/.

Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., and Gao, J. (2021). Deep learning–based text classification: A comprehensive review. *ACM Comput. Surv.*, 54(3).

MIT (1997). Mit - support vector machines: Training and applications. https://dspace.mit.edu/handle/1721.1/7290.

MIT (2018). Mit - the risk of machine-learning bias (and how to prevent it). https://sloanreview.mit.edu/article/the-risk-of-machine-learning-bias-and-how-to-prevent-it/.

MIT (2022). Mit - data warehouse at mit: Strategy document. https://ist.mit.edu/sites/default/files/services/business/Data\%20Warehouse\%20@\%20MIT_\%20Strategy\%20Document.pdf.

Mukhopadhyay, D., Sharma, M., Joshi, G., Pagare, T., and Palwe, A. (2013). Experience of developing a meta-semantic search engine. In *2013 International Confer-*

ence on Cloud Ubiquitous Computing Emerging Technologies, pages 167–171.

Nielsen (2012). Nielsen - consumer trust in online, social and mobile advertising grows. https://www.nielsen.com/us/en/insights/article/2012/consumer-trust-in-online-social-and-mobile-advertising-grows/.

Oracle (2022). Oracle - what is a data lake? https://www.oracle.com/ch-de/big-data/what-is-data-lake/.

Picard, O., Lord, M., Massé, A. B., Marcotte, O., Lopes, M., and Harnad, S. (2013). Hidden structure and function in the lexicon. *CoRR*, abs/1308.2428.

Reddy, V. (2018). Svm_sentiment_analysis repository. https://raw.githubusercontent.com/Vasistareddy/sentiment_analysis/master/data/train.csv.

SAS (2022). Sas - big data, what it is and why it matters. https://www.sas.com/en_br/insights/big-data/what-is-big-data.html.

Soltani, S., Seno, S. A. H., and Budiarto, R. (2021). Developing software signature search engines using paragraph vector model: A triage approach for digital forensics. *IEEE Access*, 9:55814–55832.

University, N. (2017). Norwich university - 5 steps for conducting computer forensics investigations. https://online.norwich.edu/academic-programs/resources/5-steps-for-conducting-computer-forensics-investigations.