# Semantic Tools for Forensics: Towards Finding Evidence in Short Messages

Michael Spranger, Eric Zuchantke and Dirk Labudde
University of Applied Sciences Mittweida
Mittweida, Germany
Email: {*name.surname*}@hs-mittweida.de

***Abstract*—Mobile devices are a popular means for planning, appointing and conducting criminal offences. In particular, short messages (SMS) and chats often contain evidential information. Due to the terms of their use, these types of messages are fundamentally different from other forms of written communication in terms of their grammatical and syntactic structure. Due to the low price of media storage, messages are rarely deleted. On one hand, this fact is quite positive as possible evidential information is not lost. On the other hand, considering only SMSs, 15,000 and more on only one cell phone is not uncommon. In the most cases of organized or gang crime, there is not one but many devices in use. Analysing this huge amount of messages manually is time consuming and therefore not economically justifiable in the cases of small and medium crimes. In this work, we propose a process chain that enables to decrease the analysis and evaluation time dramatically by reducing the messages which need to be examined manually.**

***Keywords–forensic; short message; German; text processing***

## I. INTRODUCTION

Investigations in criminal cases involve more and more investigating computers, smart phones, tablets and other devices of modern communication. This trend applies not only to computer crime in the strict sense, but also to many cases of classical crime. This is true because, on one hand, victims are easier to find and to spy on in a networked world, and on the other hand, the communications via the Internet or mobile devices have become a standard for our society and hence for sub-societies and criminals. Some of the traces of a crime, in particular those of a textual nature, are accumulated more than ever in the storage of mobile devices. Criminals use modern means of communication to plan their activities or arrange cooperatively committed crimes, as well as to find and contact potential victims. Due to the low price of media storage, messages are rarely deleted. This fact is quite positive since possible evidential information is probably not lost. However, if we consider only SMSs, experience has shown that 15,000 and more on only one smart phone is not uncommon. In addition, mobile devices may contain messages from messenger like ”WhatsApp” whose volume often exceeds ten times the volume of SMSs. If we consider gang or organized crime in general we need to realise that there is not one but many devices in use. Nowadays, the analysis of this huge amount of messages is mainly done by hand using much intuition and experience to separate the significant texts. This manual task is very time consuming and therefore not economically justifiable in the cases of small and medium crimes. Analysing and evaluating forensic texts in an automatic way is generally challenging, as shown by the authors in previous work [1] [2]. Forensic texts, as considered in this work, are texts that are subject to legal considerations with the goal of taking evidence. The analysis of such texts is regularly a branch of general linguistics [3]. In order to perform such analyses on a large amount of texts, methods from the field of computer linguistics are required. These are originated in the crossover of linguistics and computer sciences [4].

Currently, our cooperating local criminal investigation department uses Cellebrite's *Physical Analyzer* [5] for reading data from mobile devices. As a result, a multi-paged Excel or XML-based report with all the raw data reconstructed and gathered from the examined device is generated. Even if the extracted data are presented in a structured way, this particularly does not apply to the contained textual data. These remain in their original form and need to be analysed manually. If this process should be supported by automation, the special characteristics of SMSs as considered in the next section must be taken into account. There are few works dealing with the processing of SMSs. For example, Cooper et. al [6] extracted information from SMSs using manual created structural patterns in order to enrich a library database with current information. Mangan [7] has introduced an approach using structural patterns as well, but generating them by analysing the interdependency distance between slots and keywords. Amailef et al. [8] presented a mobile-based emergency response system for intelligent m-government services based on ontologies. They used a maximum entropy model for extraction of event entities from SMSs. However, we show in Section II-B, that none of these approaches is actually suitable to extract evidential information from forensic texts.

Subsequently, we propose a process chain based on these insights that enables the criminalists to reduce their search space for evidential messages significantly and hence the amount of messages that need to be analysed manually. The proposed process chain is based on an automatic clustering of coherent messages and uses a dictionary and a bag-of-words model for calculating the significance of each cluster with respect to the area of crime under consideration. In this way, the most time-consuming part in analysing SMSs can be accelerated dramatically.

In Section II, the SMS corpus we used is presented. Further, we will introduce the characteristics of this text type we found through the manual analysis of this corpus. Subsequently, in Section III the methodology used for clustering and ranking the messages are described. In Section IV some preliminary results are presented in order to give a first impression of the performance of the presented methods. After a short

 ISBN: 978-1-61208-364-3

summary in Section V, we envision some approaches for further development in Section VI.

## II. Subject of Study

### A. Forensic Short Messages Corpus

Due to a cooperation agreement between the author's university, the local criminal investigation department and the local public prosecution department, a first dataset of two closed cases of drug crime is provided. In each case, one single smart phone of the suspect has been seized and a physical backup has been created using Cellebrite's *Physical Analyzer* [5]. The backup is provided as an Excel report containing all textual data and meta-data including references to binary files from the cell phone under examination. Table I shows the amount of data currently available for evaluation. Unfortunately, only the SMSs from the first device are manually marked as evidentiary or not. In order to evaluate the results, in this work, we consider only the SMSs as well as some contact information concerning the sender and receiver of such messages contained in the report of the *HTC Desire A9191*.

TABLE I. Corpus containing SMSs

| Device | SMSs | Chat messages |
|---|---|---|
| HTC Desire A9191 | **14,307** | 132,345 |
| P743T Skate | 810 | 13,749 |

### B. Characteristics of Forensic Short Messages

Forensic text in general refers to every textual data that may contain evidential information. Their structure and quality regarding grammar, syntax and wording strongly depends on the area of the crime committed by the offenders, their level of education and their social environment. A more detailed description of the general characteristics of forensic texts can be found in [2]. Personalized SMS form the extreme case of these characteristics. They are particularly marked by frequent lack of correct grammatical structures. Therefore it is difficult to use (lexico)-syntactic pattern as in [6] [9] for extracting information of criminalistic relevance. Further, the usage of non-standardised emoticons, abbreviations, emotionally intended character extensions and especially written effects of language erosion caused by language-economic processes make this task more difficult and lead to a failing of known techniques. The following list shows some example texts to illustrate the problem:

- *"aber was ich mein[e] is[t] wir müss[e]n wenn wir weihnacht[e]n gefeiert hab[e]n* ***übelst money hab[e]n****"*
- *"Beruhig[e] dich* ***ich zieh[e] denn*** *das nächste ma[l] rich[tig]* ***fette ab****! :)))))"*
- *"Ich schreib jetzt wegen dir hab ich mein 12g nicht bekommen Weil Du* ***ne*** *aus[ de]m* ***knick*** *gekommen bist XD"*

Missing characters are included in square brackets, whereas additional characters are marked by strike-through. Slang-afflicted words and phrases are printed in bold. The most challenging problem in the considered context of SMS with criminalistic relevance is the usage of slang-afflicted language combined with terms of hidden semantics. Hidden semantics refers to one kind of a steganographic code. Such a term is used in its common innocent meaning but its actual semantic background is prearranged by a narrow circle of insiders. For example, the question

*"Bringst du ein Wernesgrüner mit? (Can you bring a Wernesgrüner?)"*

appears innocent because the term *Wernesgrüner* is used as a beer brand. But, within the actual context the meaning of this term is marijuana. Note that in this example we intentionally do not use slang to avoid misunderstandings. But commonly terms of slang are mixed in regularly. These characteristics make it difficult even for criminalists and linguists with years of experience to read and understand the semantics of forensic SMSs.

If it becomes clear that any information not found by the system may be crucial in proving the guilt or innocence of a criminal suspect, then it follows that decisions concerning the evidential value of forensic SMSs cannot be made by a machine.

## III. Methodology

In the last section, we explained why a fully automated solution should be rejected currently. For this reason, the way we propose is to decrease the effort of a manual search by reducing the search space automatically. This way the criminalist is able to find evidentiary information in a significantly shorter time.

Therefore, in this work, we outline a process chain towards reducing the search space by filtering the contacts which have exchanged significant messages and providing the corresponding conversation. The process is divided in two steps:

1) clustering of coherent messages to conversations
2) calculating a significance value for ranking conversations

### A. Clustering Coherent Messages

As we stated earlier, we cannot be sure to identify all of the significant exchanged messages contained in a corpus. But, we can increase this probability simply by trying to detect significant conversations instead of concrete messages. A conversation is by definition a set of semantically and temporally coherent messages.

More formally, let $c = m_0, ..., m_n$ whereby $c$ is a conversation and $m \in M$ a concrete message from the set of all messages in a temporal relationship contained in the corpus under consideration. In order to create clusters of coherent messages we analysed the length of the pauses between two messages. Direct comparison of these values with the areas, manually marked as significant by criminalists, reveals that long pauses are rarely situated within these areas. Figure 1 shows a clipped part of such a diagram regarding the conversation between two persons. The x-axis represents all messages of the considered subset of the corpus. The y-axis in positive direction represents the pause length between one

 ISBN: 978-1-61208-364-3 

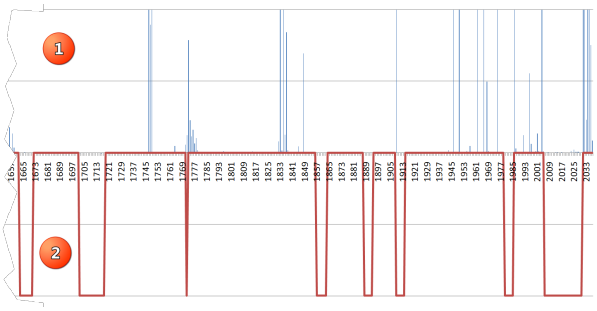


Figure 1. Clipped part of a diagram, that directly compares the pause lengths (1) of all SMS in the corpus and the manually marked significant conversation areas (2).

message to the corresponding answer. In negative direction, the y-axis shows the manually marked significant areas.

This observation leads to the approach to use the $Q_{0.75}$-Quantile of the length of pauses as threshold for the decision whether a message belongs to one conversation or is already part of the following. Applying this approach to a subset of the corpus with 3152 messages exchanged by two persons within one year, 352 conversations could be detected. The threshold was determined empirically. Experience shows that the pause length strongly depends on the individual communication behaviour. Therefore, the universality must be tested on other corpora, which, however, are currently not available.

### B. Ranking Conversations

When the set of conversations $C = \{c_0, ..., c_n\}$ have been created the next step is to find out which of these are significant regarding the object of investigation. Respecting the insights from Section II-B, we decide to use a bag-of-words model combined with a domain specific dictionary $d$ to assign a significance value to each conversation and hence to each person being part of it. This significance value $S$ can be calculated depending on the frequency of domain-specific terms (see (1)).

$$S_i = bag(c_i, d), \forall c \in C \quad (1)$$

These values form the basis of a heat scale we use to colour the contacts in the contact network established using the report data. Figure 2 shows the overall process. The starting point is a contact network based on the data gathered by the *Physical Analyzer* [5]. The exchanged coherent messages are clustered into conversations as mentioned earlier. Subsequently, the significance value is calculated for each of these conversations. Based on these values suspicious contacts and communications will be marked using the corresponding colours from the heat scale.

The determining factor for a good result is a good dictionary. A dictionary that comprises local language conditions, as well as terms from different categories of offences, is currently not available (at least in Germany). Therefore, we need to create an appropriate dictionary for each offence category and each local cultural circle before we can start to calculate the significance of a conversation.

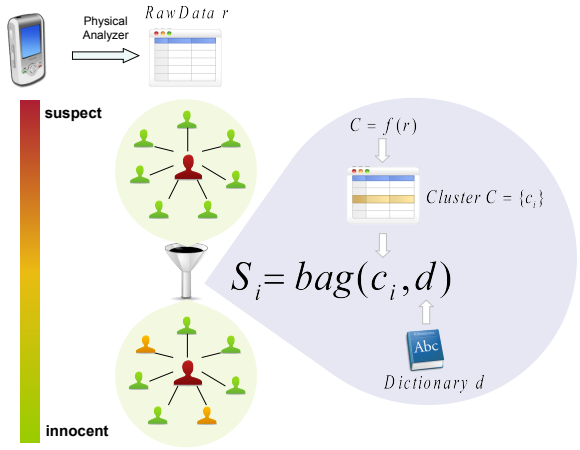


Figure 2. The process of detecting suspicious communication.

### C. Creating the Dictionary

We started dividing the corpus into significant and non-suspicious parts and performing a discriminant analysis involving stop-word elimination and stemming. Considering only the frequency classes 1 and 2 (words exclusively in suspicious texts and words relatively more frequent in such texts) we found 882 ”suspicious” terms. Using these terms in turn for processing the whole dataset for evaluation we achieve 98.5% sensitivity with 100% precision. Looking at the distribution of hits, so we found that the most of them are unique. The reason for this is due to the high number of unique spellings, caused by syntactic and typographical errors as well as deliberate word extensions. However, these lists of terms can form a basis for the dictionary, especially if more than one corpus is taken into account and words are sorted out with respect to their frequency within all corpora. In addition, it is useful

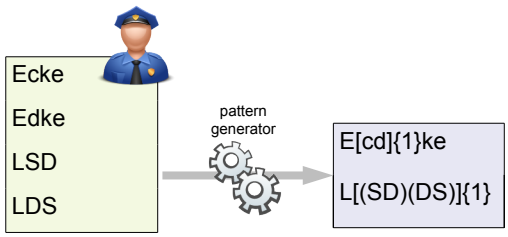


Figure 3. Generating a pattern dictionary by transforming criminalist's knowledge.

to integrate the knowledge of the local criminalists who deal with similar cases in a similar environment every day. This experiential knowledge is the best source of information for both, slang and hidden semantics. The manually added terms need to be extended automatically, for example, by twisting letters and transforming in patterns, e.g., regular expressions using an appropriate pattern generator (see Figure 3).

## IV. Performance of a Prototype

Due to the lack of other annotated corpora the first dictionary described in Section III-C has been filtered and extended manually with the help of specialists in the field of drug crime.

 ISBN: 978-1-61208-364-3

In an effort to quantify the performance of the preliminary approach described in this paper, the 352 clusters of coherent messages (see Section III-A) have been filtered using the available dictionary and employing the algorithm described in Section III-B. This implementation of the proposed process chain achieved 67% sensitivity with 100% precision. The cause of the low sensitivity is due to the coverage of the created dictionary, which is, in comparison to the coverage of a comprehensive and ready-to-use dictionary, relatively low. Therefore, the improvement of the dictionary is in the focus of further development. However, the work load necessary for manual search decreased to only 15 %.

## V. Conclusion

The manual analysis of forensic SMSs gathered from mobile devices during the criminal proceedings is very time-consuming and not economically justifiable for small and medium crimes. We have shown that extracting information from forensic SMSs in an automatic way is challenging. Considered existing methods are limited to specified domains and require a predominantly correct language usage and fail if applied on forensic SMSs. The reason for this is, among others, mainly missing standardized structures and the strong use of local dialect as well as an emotionally-influenced style of writing. Therefore, we proposed a process chain for decreasing the manual effort in analysing such messages by reducing the search space. In order to do this we cluster coherent messages to single conversations. Subsequently, we calculate a significance value using a bag-of-words model and a dictionary. Applied to a real world dataset - a closed case of drug crime provided by the local public prosecution department - we could evaluate the process chain with acceptable preliminary results.

## VI. Further Work

Currently, we are trying to improve the calculation of the significance value by applying a similar bootstrapping algorithm as presented in [2] for the field of categorising forensic texts in general.

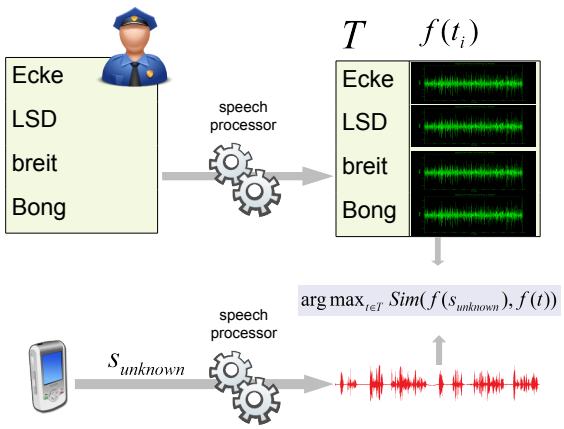


Figure 4. Dictionary containing pronunciation profiles as a basis for matching terms with high failure tolerance.

For testing the universality of the proposed process chain and especially the dictionary we need further corpora. Fortunately, the local prosecutor's office has announced the release of additional data. Another approach we currently consider is to create a dictionary, as well as a corresponding algorithm for calculating the significance value with a high failure tolerance as shown in Figure 4. Here, pronunciation profiles are used as a basis for understanding special terms.

## Acknowledgment

The authors would like to thank the members of the criminal investigation department and the public prosecution department Chemnitz (Germany). We acknowledge funding by "Europäischer Sozialfonds" (ESF), the Free State of Saxony and the University of Applied Sciences Mittweida.

## References

[1] M. Spranger, S. Schildbach, F. Heinke, S. Grunert, and D. Labudde, "Semantic tools for forensics: A highly adaptable framework," in Proc. 2nd. International Conference on Advances in Information Management and Mining, IARIA. ThinkMind Library, 2012, p. 27 to 31.

[2] M. Spranger and D. Labudde, "Semantic tools for forensics: Aproaches in forensic text analysis," in Proc. 3rd. International Conference on Advances in Information Management and Mining, IARIA. ThinkMind Library, 2013, p. 97 to 100.

[3] H. Kniffka, Working in Language and Law. A German perspective. Palgrave, 2007.

[4] K.-U. Carstensen, C. Ebert, C. Ebert, S. Jekat, R. Klabunde, and H. Langer, Computerlinguistik und Sprachtechnologie - Eine Einführung, 3rd ed. Spektrum Akademischer Verlag, 2010.

[5] C. M. S. LTD". Ufed physical analyzer - mobile daten ermitteln, dekodieren und bereitstellen. [Online]. Available: http://www.cellebrite.com/de/mobile-forensics/products/applications/ufed-physical-analyzer (2014.05.21)

[6] R. Cooper and S. Ali, "Extracting data from short messages," in Natural Language Processing and Information Systems, LNCS 3513. LNCS, Springer, 2005, pp. 388–391.

[7] D. H. W. Dannis Muhammad Mangan, "Information extraction from short text message in bahasa indonesia for electronics," Jurnal Sarjana ITB bidang Teknik Elektro dan Informatika, vol. 1, 2012, pp. 29–32.

[8] K. Amailef and J. Lu, "Mobile-based emergency response system using ontology-supported information extraction," in Handbook on Decision Making, ser. Intelligent Systems Reference Library, J. Lu, L. Jain, and G. Zhang, Eds. Springer Berlin Heidelberg, 2012, vol. 33, pp. 429–449.

[9] E. Riloff, "Automatically constructing a dictionary for information extraction tasks," in Proceedings of the Eleventh National Conference on Artificial Intelligence, ser. AAAI'93. AAAI Press, 1993, pp. 811–816.

 ISBN: 978-1-61208-364-3