MoNA: Automated Identification of Evidence in Forensic Short Messages

Michael Spranger*, Florian Heinke, Luisa Appelt, Marcus Puder and Dirk Labudde
Department of Applied Computer Sciences & Biosciences
University of Applied Sciences Mittweida
Mittweida, Germany
Email: {*spranger**, *florian.heinke, luisa.appelt, marcus.puder, labudde*}@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 potentially evidential information is not lost. On the other hand, considering only SMSs, 15,000 and more stored only on one mobile phone is not uncommon. In most cases of organized or gang crime, there is not one but many devices in use. Analysing this large 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 amount of messages, that need to be examined manually. We further present an implemented prototype (MoNA, mobile network analyzer) and demonstrate its performance.**

Keywords–forensic; ontology; German; text processing; expert system; text analysis; short messages

I. Introduction

With our previous work [1] we try to close the gap between backup and recovery of data and its content analysis in the context of mobile forensics. The fast-growing mobile market, constantly emerging or rapidly changing technologies and high hardware diversity require rapid development of new forensic tools. In recent years, many works have dealt with the backup and recovery of data on a variety of platforms [2]. However, there are few works that deal with the analysis of the textual content.

Over the last decade, our understanding of communication and its means have changed drastically. With the introduction of inexpensive messaging technologies and comfortable usability driven by increasingly powerful smart devices, communication has shifted towards conversing via chats and short messages (especially short messaging service, SMS). Besides rising computational power, mobile devices are also provided with significant amounts of memory that allow storing application data, documents, images, and thousands of messages exchanged with a multitude of conversation partners. Although the number of exchanged messages can be in the thousands, they account only for a small fraction of occupied space in general. In consequence, chat and SMS logs are rarely deleted.

In the context of digital forensics, this aspect leads to an ambivalent situation. On the positive side, it has become more likely that confiscated devices yield information relevant for the investigation process and could reveal additional evidential aspects, such as identities of backers or other crime-related intentions of the suspect. On the downside, these information need to be extracted from the raw chat data, which is, considering its scale, barely manageable by manual means. In addition, with the growing amount of available memory and the ongoing popularity growth of text messaging, it can be postulated that manual perusing and annotating will become practically impossible. Hence, there is a necessity in developing computational (semi-) automated technologies that can support the investigator in the process. To achieve this, researches have to cope with a number of issues. Most notably, messages are often enriched with typos and grammar mistakes introduced by lowered language use standards observable in casual text conversations. Such mistakes pose major problems for text mining and computer linguistics.

Based on our previous work [1] a straightforward technique for identifying individual conversations in SMS chat logs is introduced in this paper and evaluated on a manually-annotated SMS dataset. In addition, in Section II we first discuss general background and related work considered in the MoNA development process. In Section III, we define and characterise short message semantic analyses in the context of forensics, providing detailed aspects involved in the motivation of our work. Further, the SMS dataset used to develop and evaluate MoNA is described. We emphasize that the dataset in use has been relevant for a closed drug crime investigation, thus actual information provided in this study are based on a real-life application scenario. In this respect, in Section IV we introduce measures for quantifying the potency of a keyword dictionary provided by investigators that is used by MoNA to classify and score identified conversations. These measures additionally provide statistical figures that can be used to further optimise and refine the dictionary in use. After discussing the evidential conversation detection process utilised by MoNA (Section V), we demonstrate its performance in Section VI and provide future prospects in Section VII.

II. Related Work

Compared with texts from industry, medicine or science, relatively few works deal with the analysis of short messages. Most of these works address the binary classification problem in terms of SPAM detection. For example, in a large-scale study Skudlark [3] examined approaches to detect SPAM activities. However, they rely on the presence of URLs in the text body, which limits the applicability of these methods to forensic short messages on very few cases of fraud, computer sabotage or similar crimes. Ahmed *et al.* [4] presented an SMS classification approach based on Naive Bayes and *a priori* algorithms. A further method has been discussed by Xu *et al.* [5] that relies on content features for classification. Although this method yields reasonable results, an application to most fields of forensics cannot only be based on meta data. But this kind of data can be useful to enlarge the target matching space. In the field of multi class SMS classification Al-Talib *et al.* [6] introduced a technique using an improved TF-IDF weighting, whereas Patel *et al.* exploit artificial neural networks [7]. Another interesting work was presented in 2011 by Ishihara [8]. In this work, the author proposed a likelihood ratio-based approach for SMS authorship classification using n-grams. The model was trained and evaluated using the NUS SMS corpus [9]. Unfortunately, a similar corpus for German forensic SMS is currently not available. The general problem when trying to create such a corpus in the field of forensics is the availability of real-life data. This is the reason why Ishihara considered non-forensic data, while the developed classifier has been trained and evaluated for forensic purposes. Therefore, with respect to forensic short messages and their special characteristics (which are discussed in Section III-B in more detail), the applicability of such a classifier and its performance in the real-life context of forensics remains unsettled. Furthermore, approaches for extracting information from short messages, as discussed in [10], [11], are frequently based on the presence of correct grammatical structures. However, these do not exist in most cases for short messages. Knowledge-based approaches, such as proposed in [12], are more promising since they can include *a priori* knowledge of the investigator to support information retrieval as well as information extraction processes. In the work presented by Nebhi [12] Twitter posts are considered, which, however, are similar to forensic SMS in some degree.

Thus, in general, there is no approach available that can cope with the challenges posed by real-life short message data. In addition, such an approach is required to be both applicable to all the data, and to perform as reliable as required by forensic investigation standards.

III. Background

A. Forensic Short Messages

The analysis of short messages is a particular challenge in the context of forensic text analysis. The reasons for this is the combination of forensic text characteristics and of high information density, which is characterized by limitations in the number of characters. Such limitations arise on the one hand by technical reasons, on the other hand by the kind of use. Thus, short messages are often used in terms of “by the way messages”.

Definition 1 (Forensic Short Message): In this study, any textual message having the properties of an incriminated text and is sent or received using a short message service is considered a forensic short message.

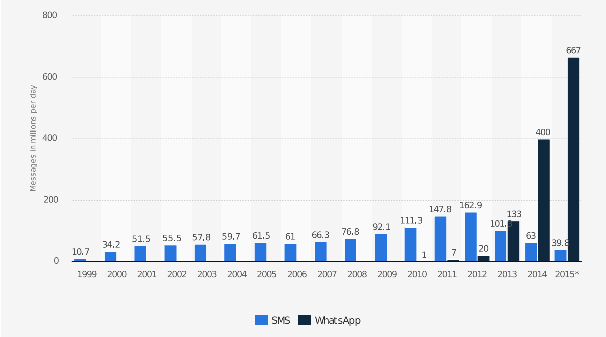


Figure 1: Number of SMS and IM messages (WhatsApp) sent in Germany from 1990 to 2014 in millions per day (2015 estimated) [13].

Looking at current surveys of the short message traffic in the Federal Republic of Germany (see Fig. 1), a turnaround can be seen in the development starting in 2013. The reason for this is not a decrease in the mobile communications in general, but in a shift to other convenient communication services such as instant messaging services (e. g. Whats App). Nevertheless, every citizen sends one instant message per day on average. The outstanding role of text messages today and in the future, both in general and forensics contexts, was thoroughly discussed in [14]. Since the communication behaviour is mainly influenced by the type of use, the change of the medium has had only a relatively small impact on the writing behaviour of the user. Thus, the results presented in this work can be transferred in principle to other forms of mobile communication in writing. In general, the forensic analysis of incriminated texts is a big challenge for investigators—which is especially the case for short messages. In addition to large message quantities only sent by one individual, such messages have a particular characteristic, which makes the analysis difficult even for experienced investigators. Considering the amount of content that needs to be fully read, this effort is probably justifiable only in cases of severe crime or crime of high public interest. One of the current biggest limitations during the development of an automated solution for forensic purposes is the lack of a gold standard. Yet, an effective and efficient analysis in each relevant case is unthinkable without the usage of computational solutions.

B. Characteristics of Forensic Short Messages

As already discussed in [1], forensic text’s 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 [15]. 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 discussed in [16] [17] 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] den~~n~~*** *das nächste ma[l] rich[tig]* ***fett~~e~~ 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 shown as strike-through text. Incorrect capitalisations are underlined. 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 refer 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 and unsuspicious because the term *Wernesgrüner* (a German beer brand) is used as in asking for a bottle of beer. However, by considering the actual context, the author of this message is actually asking for 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.

Thus, it becomes clear that any information not identified as relevant by an automated system may be crucial in proving the guilt or innocence of a criminal suspect. Eventually, it can be stated that decisions concerning the evidential value of forensic SMSs cannot be made by a machine.

C. Dataset under Consideration

The data used by the authors for the development of MoNA is based on a dataset of a closed case of drug trafficking provided by a cooperating prosecutor for research purposes. Nevertheless, the data is not publicly available. For this purpose the legal framework has to be established, at least in Germany. In the case under consideration a smart phone of the suspect, an HTC Desire A9191, has been seized and a physical image has been generated by using Cellebrite's *UFED Physical Analyzer*. The textual data contained in this image was exported as an Excel Workbook and forms the basis for all further investigations. This dataset includes 14,307 short messages (SMS) and 132,345 chat messages. During the development of MoNA, only SMS messages have been used so far. Through an official of a cooperating police department all short messages were manually read and evidential ones were labelled as relevant. Afterwards, the same work was performed by a member of the research team without criminalistic background.

In summary, only half of the relevant messages were correctly classified as evidential by the research member and, on the contrary, messages considered as insignificant by the investigators were classified as evidential. This shows that subjectivity can introduce significant errors in analyses processes and emphasises the need for expert knowledge. This study thus focuses on the prototypic implementation of MoNA as a strategy for identification and classification of conversations with respect to their relevance to the crime in question.

IV. Word Dictionary Potency

The majority of text mining and computer linguistic algorithms rely on word dictionaries that provide the initial set of words, which are screened against a given text dataset in the process. In computer forensics, the investigator aims at maximizing the number of identified messages containing evidential information, to which we refer to as significant messages in the following text. In general, the basis for the successful identification of significant messages or conversations in large message sets using string matching techniques and phonetic algorithms predominantly requires a potent word dictionary. Word dictionaries are subjected to two major requirements: First, they have a significant impact on classification performances of utilised methods and thus should be optimally composed in this respect. Secondly, word dictionaries are required to be domain-specific, case-independent, and generalized corpora of words—meaning that each should be interchangeable and not be specifically tailored toward the dataset in question. Especially in the field of computer forensics, it needs to be further emphasized that a dictionary is considered to be specific for a certain time period and region as well.

In this study, MoNA has been provided with a dictionary of 90 words specific for the drug scene currently present in the Chemnitz/west Saxony region of Germany. In this section, measures of dictionary potency, which supply simple quantifications of per-word performance, are introduced, demonstrated and discussed on the available data. First, measuring initial classification power of dictionary words is demonstrated. Subsequently, it is shown how word heterogeneity of obtained word matches in the dataset can be measured. Word match heterogeneity provides statistical figures on word diversity in and between matching word sets in significant and insignificant (non-relevant) messages, which in turn can be used to represent per-word specificity. Here, the investigator aims at maximizing diversity between word sets and reduce heterogeneity within the sets, thus reducing ambiguities of obtained matches.

A. Overview on Individual Dictionary Word Potency

Analyses of dictionary potency have been conducted by employing string matching and two phonetic algorithms (Kölner phonetic [18] and Double Metaphone [19]) on the provided SMS dataset. All three algorithms reported a total of 11,665 matches in the dataset, of which 310 had been annotated as significant by the investigators. Interestingly, a large fraction of 42 dictionary words matches exclusively to either significant or insignificant messages (see Table I). The seven words only matching to significant messages account for eighteen of the 310 significant matches. Furthermore, 35 dictionary words yield no significant matches in the dataset, classifying a total of 17% of the matches as insignificant.

In theory, a powerful dictionary yields significant matches only. However, the statistics obtained from actual data show

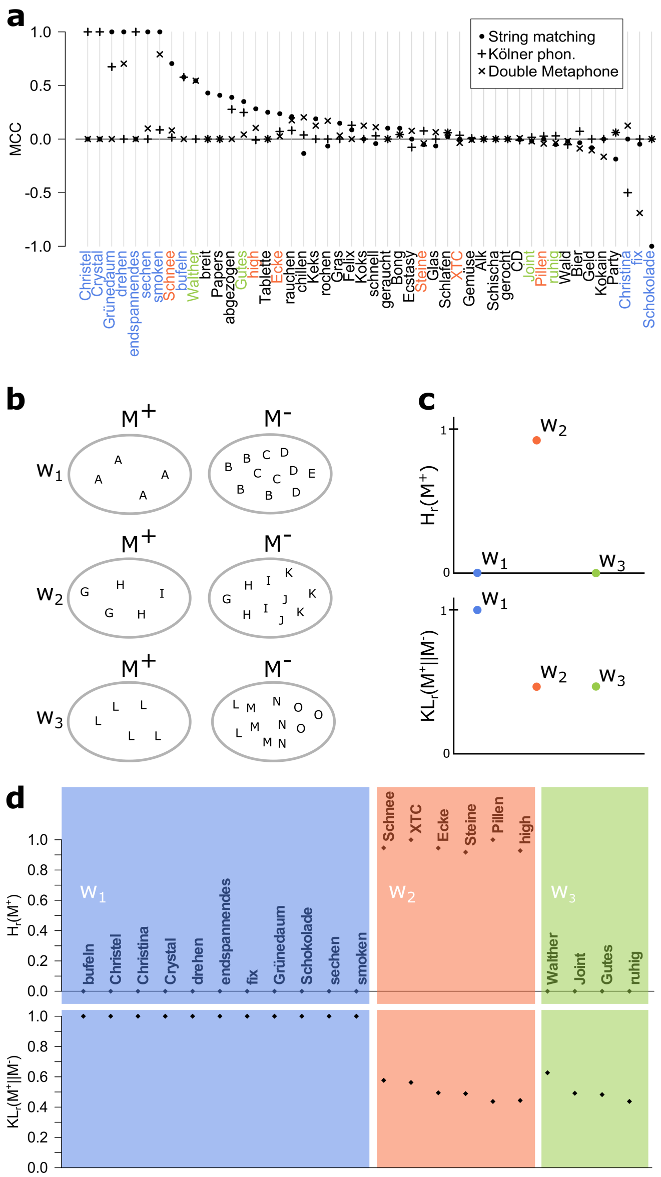


Figure 2: **a:** Predictive performance evaluated by means of Matthews correlation coefficient (MCC) of 90 dictionary words yielding matches in significant and insignificant messages (see Table I). **b** and **c:** Relative entropy (H_r) and relative Kullback-Leibler divergence (KL_r) are measures for assessing word match set homogeneity respectively set divergence. Three illustrative word match homogeneity scenarios for words w_1, w_2, and w_3 are depicted, with schematic plots of H_r and KL_r obtained from these scenarios shown on the right (M^+ and M^-: matching word sets of significant respectively insignificant messages). **d:** Plots of H_r and KL_r of 21 words yielding $KL_r > 0.4$. Colour highlighting is in correspondence to the three word matching scenarios shown in **b** and **c**, indicating varying degrees of ambiguity and, thus, significance to dictionary power.

mixed power of individual dictionary words. The aspect that thirty-five of 90 words are anti-correlated—yielding only insignificant matches—is rather surprising, as one would expect most of the matches to be exclusively significant or at least to be matching to both cases. Although errors should be expected in practice, large fractions of anti-correlated words, as observed in this study, highlight that a given domain-specific dictionary can produce unwanted effects caused by hidden ambiguities in the data. Hence, a dictionary should always be considered as a set of independent words—each of these with its own meaning and power with respect to classification performance.

TABLE I: Results of initial word dictionary testing. 90 domain-specific dictionary words have been matched against the SMS dataset using string matching, Kölner phonetic and Double Metaphone. If a matching message contains evidential information, the match is considered significant.

number of dictionary words	*type of matches*	*% of all matches*
7	only in significant messages	0.2
35	only in insignificant messages	17.0
48	both	82.8

B. *Measuring per-word Classification Power*

To demonstrate the varying power of words, classification performances of the 48 words matching to significant and insignificant messages have been analysed. Here, the Matthews correlation coefficients (MCC) [20] have been computed for each word and each of the three used algorithms. The MCC is defined as follows:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}. \quad (1)$$

The classification statistics TP (true positives), FP (false positives), TN (true negatives), FN (false positives) correspond in this context to the following numbers:

- TP - number of reported matches corresponding to significant messages
- FP - number of reported matches corresponding to insignificant messages
- TN - number of cases where the algorithm indicates no match in insignificant messages
- FN - number of cases where the algorithm indicates no match in significant messages

The MCC is in the range of -1 and +1, where +1 corresponds to perfect classification, whereas the number of FN and FP cases is 0. In contrast an anti-correlated performance is indicated by an MCC of -1. Furthermore, although the MCC is a more strict measure of classification performance in comparison to the classic F1-measure, it is less prone to errors introduced by class imbalance, which is present for the majority of investigated words. The MCC thus supplies an intuitive representation of dictionary potency, but also provides a measure that is ought to be maximized. The computed individual MCC values are shown in Fig. 2a. As depicted, ten words yield MCC values > 0.5 for any of the three methods. These words include the two phonetically similar words *Christel* (a German nickname) and *Crystal*, as well as *endspannendes* (misspelled German word for 'anything that is relaxing'). In case of these three words, phonetic algorithms reported matches highly correlated to significant messages, however, no string matches have been identified. This is simply due to spelling mistakes made by the conversation partners. Here, phonetic algorithms have been able to successfully identify correspondences missed by string matching that had been proven to be case-relevant. Further, in case of *Schokolade*, a domain-specific synonym for hashish, anti-correlation is observed, yielding string matches only to insignificant messages. Manual inspection of string and phonetic matches revealed spelling differences between significant and insignificant messages, where *Schokolade* is written correctly in the latter, leading to reported anti-correlated string matches. In significant messages, however, a consistent miss-spelling, *Schockolade* instead of *Schokolade*, is present. Although there are only two cases of positive messages, it can be proposed that this typing error is actually made on purpose—where the additional 'c' could abbreviate 'cannabis' and therefore might encode the actual meaning of the message in addition to using the synonym. In summary, for the majority of these 48 words classification performance is insufficiently low. Even though string matching seems to perform superior to the utilised phonetic algorithms in this study, no significant performance difference is observable (one-sided Welch test, p = 0.24 for Kölner phonetic, p = 0.05 for Double metaphone).

C. Measuring per-word Match Ambiguity

Finally, if a dictionary word results to two sets of matching words in significant and insignificant messages (as in case of forty-eight dictionary words in this study), it is at least desirable to obtain two divergent, homogeneous sets of matches. Divergence indicates that the two sets differ in matching word composition, whereas homogeneity indicates word diversity within the sets. Thus, both measures in combination can provide quantifications of ambiguities present between and within both sets. In order to avoid ambiguity, the researcher aims at increasing divergence and homogeneity. Subfigures 2b and c illustrate resulting hypothetical word match scenarios for three imaginary dictionary words (w_1, w_2, w_3). M^+ and M^- correspond to the sets of matching words in significant respectively insignificant messages. For measuring word set divergence and homogeneity, relative Kullback-Leibler divergence (KL_r) and relative Shannon entropy (H_r) are here proposed. In general, KL_r corresponds to the normalized Kullback-Leibler divergence, which is used to measure the difference between probability distributions P and Q. A probability distribution for a word set M can be simply deduced by considering the relative frequency of each word in the set of unique words U_M derived from M. Thus

$$P(w \in U_M) = \frac{f(w)}{|M|} \tag{2}$$

is obtained. For clarity, $P(M^-)$ is denoted as $Q(M^-)$ in the following. From this, the Kullback-Leibler divergence between sets M^+ and M^- can be readily computed:

$$KL(M^+||M^-) = \sum_{w \in U_M} P(w) \log_2 \frac{P(w)}{Q(w)}, \tag{3}$$

where $M = M^+ \cup M^-$. However, divergence as defined here results to a non-symmetric measure, which also not always strictly considers all unique words in M. In case of a given word w being only present in M^+, $Q(w) = 0$ and the Kullback-Leibler divergence is not defined. Due to these drawbacks, the two-sided divergence is utilised in this study instead:

$$KL(M^+||M^-) = KL(M^+||M) + KL(M^-||M). \tag{4}$$

Finally, KL_r can be computed by normalizing the observed divergence using the theoretical maximal divergence

$$KL_{max}(M^+||M^-) = \log_2\left(\frac{|M|}{|M^+|}\right) + \log_2\left(\frac{|M|}{|M^-|}\right) \tag{5}$$

$$KL_r(M^+||M^-) = \frac{KL(M^+||M^-)}{KL_{max}(M^+||M^-)}. \tag{6}$$

Using the definition of word probability given in equation (2), Shannon entropy can be computed:

$$H(M) = -\sum_{w \in U_M} P(w) \log_2 P(w). \tag{7}$$

In order to obtain normalized comparable quantities $H_r(M)$, $H(M)$ is weighted by taking into account the maximum theoretical entropy, leading to

$$H_r(M) = \frac{H(M)}{\log_2(|U_M|)}. \tag{8}$$

In Fig. 2 KL_r and H_r are illustrated schematically for three word match scenarios w_1, w_2 and w_3 (see subfigures b and c), which are observed for 21 dictionary words. Here, only dictionary words yielding minimal ambiguity $(KL_r > 0.4)$ are considered. As illustrated, if all unique words are uniformly distributed over M^+ and M^-, H_r is observed to be 1 (maximal) and KL_r is 0 (minimal). In this case, ambiguity is maximal and, simply by considering matching words, no classification can be achieved. The corresponding dictionary word w is thus of low classification potency. This case is similar to word matching scenario w_3, whereas four dictionary words considered in this study yield similar values (highlighted in green in Fig. 2a and d). Analogously, 11 and 6 dictionary words can be assigned to scenario w_1 (blue) and w_2 (red), respectively. Also, by taking into account MCC values of these words, the dictionary words yielding good classification potency can be identified. In our study, the dictionary words *bufeln, Christel, Crystal, drehen, endspannendes, Grünedaum, sechen, smoken* yield a low degree of ambiguity and good classification performance. Furthermore, dictionary words yielding anti-correlation correspond to word matching scenario w_1 as well.

In summary, these statistics can be used to visualize individual dictionary word potency, but also to provide measures that can aid in identifying unknown correlations and selecting additional potent words from obtained matches, or replacing ambiguous less potent words. In this study, it is apparent that a majority of words provided by the dictionary are of low potency. A large fraction of words are not sensitive to significant messages or provide only low classification potency with respect to message relevance.

V. Identification of Evidential Short Messages

With respect to properties of forensic short messages discussed in Section III-B, the identification of crime-relevant messages within a large message history is a classification problem that is difficult to solve by means of computational approaches as well as manual annotation. However, by taking into account information extracted from related conversations instead of individual messages, automated classification strategies could be developed and applied providing the investigator with a list of conversations, which in turn can be manually perused, put into context, and can thus aid in the investigation process. As a positive side effect the context of the message is maintained in such a preprocessing strategy, which facilitates the understanding when manually perusing obtained classification results. Thus, an automated method for identifying individual conversations in message histories is desirable. In this section, a statistical approach is suggested, which aims at addressing this problem. First, the initial strategy for extracting statistical data from message logs is elucidated and mathematical formalisms are introduced. Furthermore, a statistical measure for quantifying conversation detection performance is introduced and applied to the proposed strategy. In this study, the conversation identification strategy is applied to two drug crime-related message histories both containing manually annotated relevant (evidential) messages, whereas one further contains a conversation index obtained from peruse. The latter message history is eventually used to measure identification performance of the proposed strategy.

A. Conversation Detection

In the context of this study, a conversation is considered as any amount of time- and semantically-coherent messages between at least two people. Formally expressed let M be the set of all messages, where $m \in M$ is corresponding to any message in M. Furthermore, M is in chronological order, creating a temporal connection between the logged messages. Therefore, the chronology of the exchange between the conversational members can be tracked. In addition, the response times (the elapsed time between two sequential messages m_i and m_{i+1}) can be derived. The strategy presented here is based only on derived response times and follows a simple hypothesis: the longer the response time between two messages m_i and m_{i+1}, the lower the likelihood that both messages belong to the same conversation. Based on this hypothesis, the following approaches may gradually lead to the proposed statistical strategy:

1) Response times follow a statistical distribution (frequency distribution). Short response times are thereby more often observed than long response times.
2) Given a sufficiently large dataset and obtained response times, a probability distribution and hence a probability density function can be estimated empirically on the basis of the observed response time frequency distribution.
3) Given any response time t, the relative number of expected response times $\leq t$ in a chat log can be estimated based on the approximated density function.
4) The reversion of the above statement leads to an approach for solving the problem and is as follows: for which response time t is a given fraction p of response times with $\leq t$ to be expected?
5) If p is chosen sufficiently small ($p = 0.05$ is a common statistical threshold), a critical response time t can be determined. Thus, it is expected that for this particular response time $95\%(1-p)$ of all messages are answered within that time period. The remainder of 5% is negligible in accordance to p.
6) If the response time between two messages m_i and m_{i+1} exceeds the critical response time, the probability for both messages belonging to the same conversation is expected to be low. Thus, it can be postulated that both messages do not belong to the same conversation.
7) M can be split into different conversations solely from the sequence of response times with respect the estimated critical response time.

With the general hypothesis elucidated, the underlying formalisms resulting to the identification strategy are now introduced. Let $\delta t = t_{m_{i+1}} - t_{m_i}$ be the response time between two sequential messages of two conversation partners. Then ΔT is the set of all response times which fall within interval $(t_1, t_2]$; thus $\Delta T = \{\delta t \mid t_1 < \delta t \leq t_2\}$. The function of all observed frequencies $h_i = \parallel \Delta T_i \parallel$ parallel over time gives a characteristic frequency distribution, which is illustrated in Fig. 3a for a message history containing 1,550 messages. The bin interval is 5 seconds.

In this histogram it can be seen that the frequency distribution follows an exponential decay of form ae^{-bt}. Therefore, short response times are frequently observed. A causal relationship between the observed distribution and response time can be explained by the responsiveness of two callers. However, this responsiveness is not constant, but varies continuously over time. To illustrate this underlying hypothesis, let t be the time that has elapsed since receipt of the last message. The recipient of the last message has not yet responded to this. So the responsiveness is B_t. It should be noted, that B_t is corresponding to a sum of several human factors, for example for this readiness, the duration of the call and already discussed aspects contribute to responsiveness variance. However, a general decrease in responsiveness can be assumed considering a general statement: if the recipient of the last message has seen no reason to answer at time t, readiness to continue the conversation does not increase at a later time $t + dt$. This responsiveness is formally expressed as B_{t+dt}. Although numerous human factors account to variance, the statement $B_t > B_{t+dt}$ is reasonable to assume in the general case. Thus, the responsiveness decreases tendentiously. Here it can be postulated that a constant reduction rate r describes the reduction of B as a function of elapsed time since receipt of the last message. This relationship can be formulated as:

$$\frac{\mathrm{d}B}{\mathrm{d}t} = -rB \tag{9}$$

Hence, the more time has passed since the receipt of a message, the lower is the probability that a response will still be sent. Equation (9) can be readily solved (equation (10)). Thus, from these considerations time-dependent responsiveness B_t results from initial standby responsiveness B_0 and constant reduction rate r. This aspect describes the exponential relation shown in

Fig. 3a.

$$B_t = B_0 \mathrm{e}^{-rt} \tag{10}$$

Furthermore, equation (10) can be understood as a probability density function obtainable by fitting parameters B_0 and r to the observed response time distribution. Optimal fitting can be determined by regression errors (see equation (11)).

$$\{r_{opt,B_{0_{opt}}}\} = \arg\min_{r,B_0} \sum_t |B_t^{calc} - B_t^{obs}| \tag{11}$$

Based on this approximated probability density function, a threshold time t_p is calculated corresponding to a sufficiently low answer probability. If this time t_p is exceeded, the conversation is considered as terminated. This probability can be assumed to be sufficiently small and is hereinafter referred to p. In this study $p = 0.05$ is observed to perform well on both considered message histories. Finally, t_p can be determined by simply integrating the approximated probability density function and corresponding rearranging of the equation:

$$\mathrm{F}(B) = \int_0^{t_p} B_0 \mathrm{e}^{-rt} \mathrm{d}t = 1 - p \tag{12}$$

The red line in Fig. 3a shows the result of the performed regression on a message history consisting of 1,550 messages. The dashed blue line illustrates the calculated t_p for which the probability of receiving a response to a sent message is lower than $p = 0.05$. For this dataset t_p is 217 seconds. In summary, proposed statistical conversation identification requires to approximate the probability density function from the conversation-characteristic response time distribution, estimate a critical threshold time t_p for a preselected p, and finally split M into disjunct sets of messages, whereas $|t_{m_i} - t_{m_{i+1}}| > t_p$.

To check if conversations containing evidential messages are erroneously split by this approach, their response time distribution has been investigated. As shown in the histogram in Fig. 3b the response time for only one relevant message exceeds the estimated t_p. Hence, all conversations containing evidential information are not falsely split in this case. However, this single message is a solitaire, meaning it is a single unanswered message without contextual references. The algorithm for detecting conversations is only applied to phone numbers at a minimum of 7 digits without any country code or at least 10 digits with country code, because shorter numbers mostly belong to telephone services and, therefore, are of less interest. Furthermore, the cut-off calculation as the first step during the conversation detection is only considered, if the number of messages between two conversation partners is at least 20. To avoid extremely short and, therefore, less meaningful conversations a cut-off value of 2 minutes is set as a minimum. These dialogs also include questions detected via the question mark symbol ("?"), as well as in combination with an exclamation mark ("!"). It was analysed whether and to which extent at these points (between question and answer/reaction) clustering occurs and if, therefore, coherent conversations were cut undesirably. The analysis results on the test dataset can be seen in Table II. Based on a manual review, it could be found out that SMSs, which contain question marks mostly (82.69 to 100%), follow up answers or reactions from the other participant and, therefore, can be treated as a coherent conversation. The cut-off value itself does not change, because the clustering/conversation detection is following the cut-off calculation.

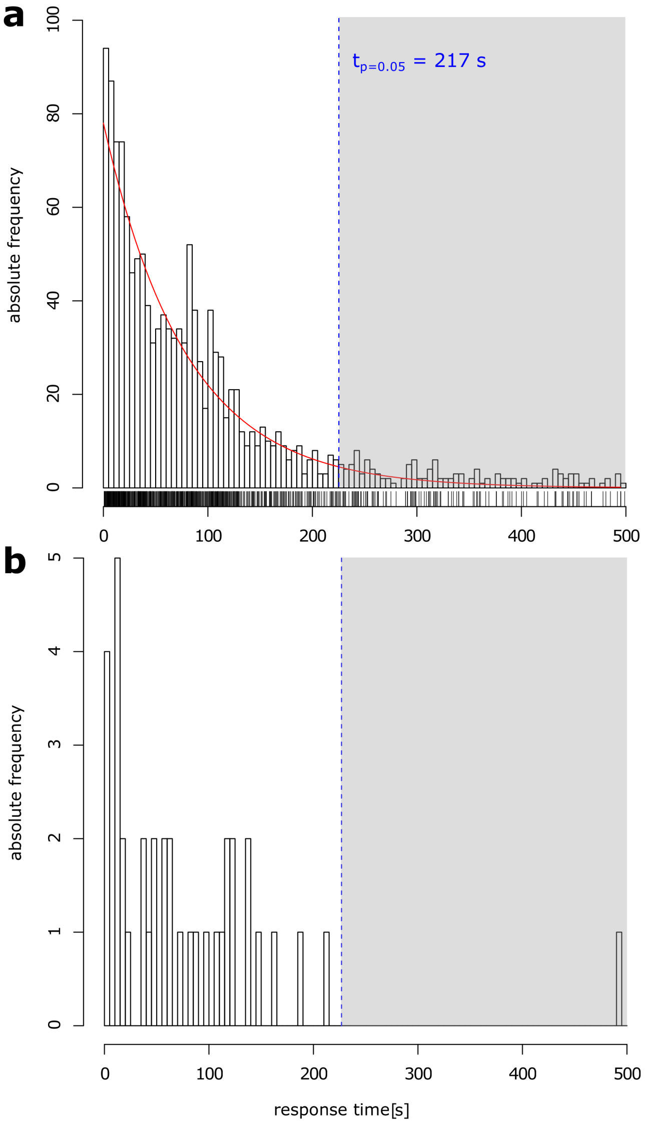


Figure 3: **a:** Response time histogram of a message history containing 1,550 messages, of which 40 were evidential for a recently completed drug crime investigation. Detecting conversations (grouping of chronologically ordered messages into disjunct sets) statistically as proposed in this study utilises a probability density function estimation (shown by red line) based on response time distribution. Employing a probability threshold p gives a critical response time t_p, upon which two consecutive messages are assigned to two separate conversations if the observed response time between said messages exceeds t_p. **b:** Frequency distribution of 40 evidential messages. As shown, utilising a t_p of 217 seconds leads to a set of conversations in which conversations containing evidential messages are not erroneously split by the proposed approach.

By reference to Table II it would appear that special treatment or filtering of question marks has a major impact on the number of clusters in the test dataset. Without question mark treatment 384 clusters arise. If the clustering is extended by the question mark treatment it results in 332 clusters

TABLE II: Comparison of the question mark consideration in the conversation detection of the test dataset.

	without "?" treatment	*with "?" treatment*
# matches	52	0
# conversations	384	332

(13.54% or alternatively 52 less clusters than before). In view of these results, the question mark treatment is recommended.

B. *Evaluation of Conversation Identification*

Evaluation of the proposed automated approach is carried out by comparing the similarity of the generated chat log partitioning with the partitioning obtained from manual annotations using an information theoretic approach. More precisely, given the two partitions normalized mutual information is computed, which indicates information coupling between both and reports the degree of uncertainty that the partitioning obtained by the proposed approach is meaningful with respect to manual annotations. A partitioning of a chat log corresponds to a set $C = c_1, \ldots, c_k$ of k identified conversations. Let $t_E(c_i)$ be the time elapsed during conversation c_i. Trivially, the sum of all t_E equals the elapsed time between all messages. According to this the ratio $p(c_i) = t_E(c_i)/\sum(t_E(c_j))$ corresponds to the probability of any randomly chosen message (or any arbitrary point of time between t_1 and t_n) to belong to conversation c_i. In this respect, a pair of conversation sets obtained by manual annotation C^{man} and automated identification C^{auto} can be compared by means of these probabilities. Considering manual conversation identification to be error-free, optimal automated conversation identification is achieved, if $C^{man} = C^{auto}$. By considering underlying probability distributions $P(C^{man})$ and $P(C^{auto})$, the statement holds analogously true if $P(C^{man}) = P(C^{auto})$. In order to evaluate the quality of the proposed approach, both $P(C^{man})$ and $P(C^{auto})$ can be utilised to measure set similarity. However, since both sets are not necessarily of equal size and there is no one-to-one correspondence between conversations in both sets, rather straightforward measures of set similarity, such as the Jaccard index or the previously discussed Kullback-Leibler divergence as well as correlation analyses of conversation indexes, have to be considered as unsuitable. Due to these constraints, normalized mutual information (NMI) is chosen for evaluation instead. In general mutual information quantifies the emitted information (or dependence) between two variables X and Y by means of scaling the joint distribution $P(X, Y)$ of both using the distribution of marginal probabilities $P(X)P(Y)$. This measure is still applicable in case when the sizes of both sets are unequal, and also does not rely on one-to-one relations. Further, mutual information can readily be normalized ($NMI \in [0, 1]$) using the marginal entropies $H(X)$ and $H(Y)$ to obtain comparable quantities. The maximum value of 1 is thus observed if the discrepancy between joint distribution and marginal distribution is maximized— which is only if $P(X) = P(Y)$. With respect to $P(C^{man})$ and $P(C^{auto})$, a maximum value of 1 is only achieved, if both sets are equal and, hence, perfect conversation identification is obtained. If C^{auto} is simply generated by chance and the identified conversation set is thus expected to be of low quality, the discrepancy between both distributions is observed to be relatively small, leading to relatively low NMI. For conversation identification NMI is defined as

$$NMI(C^{auto}, C^{man}) = \frac{2MI(C^{auto}, C^{man})}{H(C^{auto}) + H(C^{man})}, \quad (13)$$

where

$$MI(C_1, C_2) = \sum_{c_i \in C_1} \sum_{c_j \in C_2} p(c_i \cap c_j) \log_2 \left(\frac{p(c_i \cap c_j)}{p(c_i)p(c_j)} \right) \quad (14)$$

and

$$H(C) = - \sum_{c_i \in C} p(c_i) \log_2 (p(c_i)) . \quad (15)$$

$p(c_i \cap c_j)$ corresponds to the time fraction of the overlap between two conversations.

The proposed strategy was applied to a second message history of 2046 messages. The history was manually perused and 116 individual conversations were identified manually. The statistical approach utilised on this dataset yielded a $t_{p=0.05}$ of 2907 seconds. The corresponding NMI was computed and compared to NMI values resulting for critical response times t_c in the range of 30 to 30,000 seconds. This comparison provides a robustness test for the conversation identification obtained from $t_{p=0.05}$. As shown in Fig. 4 NMI$_{t_p}$ is within the response time interval (2000 s - 7000 s) that yields best performance when considering a constant critical response time as proposed here. Note that the NMI for small critical response times ($t_c < 100$ seconds) of about 0.95 corresponds to a message history-characteristic baseline performance, which is the result of marginal correlation between C^{auto} and C^{man} in this t_c range.

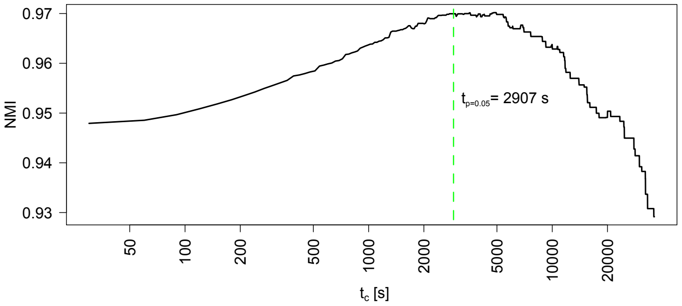


Figure 4: NMI values computed for a message history with available C^{man} determined by peruse. NMI is obtained by deriving C^{auto} for each corresponding critical response time in the range of 30 to 30,000 seconds. The critical response time $t_{p=0.05}$ computed by the proposed statistical approach (2907 seconds) is here highlighted by a green line. Conversation identification obtained by proposed approach is within the critical response time range resulting to best classification.

C. *Detecting Evidential Conversations*

Given the set of identified conversations $C = \{c_0, ..., c_n\}$, the next step is to determine which of these are significant regarding the object of investigation. With respect to the insights provided in Section III-B, we utilised 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 equation (16)).

$$S_i = bag(c_i, d), \forall c \in C \qquad (16)$$

These values form the basis of a heat scale we use to colour the contacts in the contact network established using the report data. Fig. 5 shows the overall process. The starting point is a contact network based on the data gathered by *Physical Analyzer* [21]. Exchanged coherent messages are subsequently clustered into conversations as proposed. The significance value is calculated for each of these conversations. Based on these values suspicious contacts and communications are highlighted visually on the contact network using the corresponding heat scale colours via the MoNA user interface.

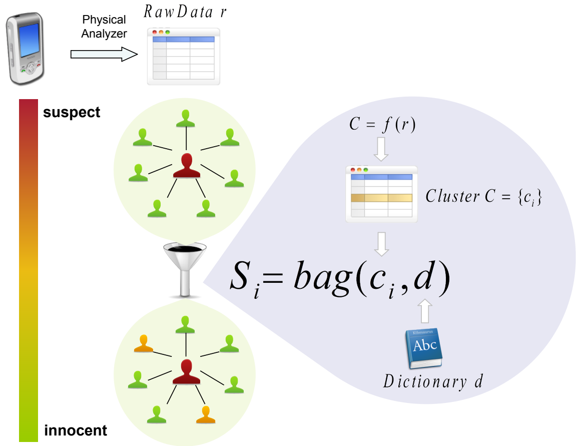


Figure 5: The process of detecting suspicious communication.

As discussed in Section IV, the determining factor for satisfactory results is a potent 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, an appropriate dictionary for each offence category and each local cultural circle is required to be created before calculating conversation significance.

D. 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 identified 882 ”suspicious” terms. Using these terms in turn for processing the whole dataset for evaluation we achieve 0.98 sensitivity with 1.0 precision. Looking at the distribution of hits, we observed 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 removed according to their frequency within all corpora.

In addition, it is useful to integrate the knowledge of 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. Such 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 Fig. 6). Current work aims at improving dictionary potency by applying a similar bootstrapping algorithm as presented in [15] for the field of categorising forensic texts in general.

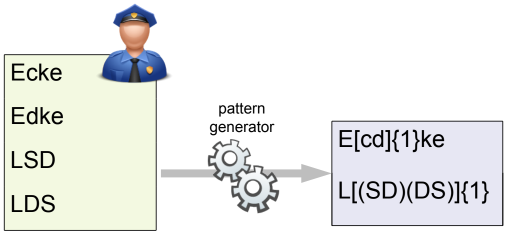


Figure 6: Generating a pattern dictionary by transforming criminalist’s knowledge.

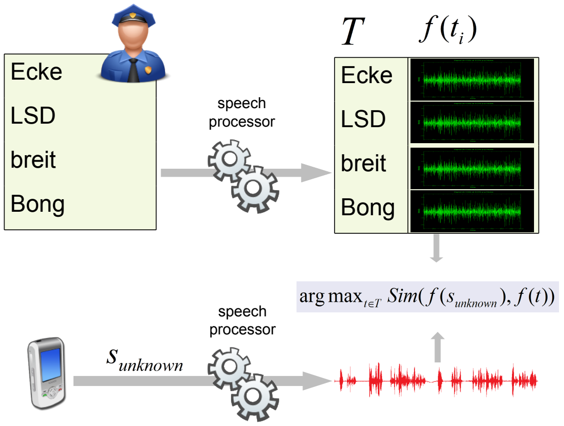
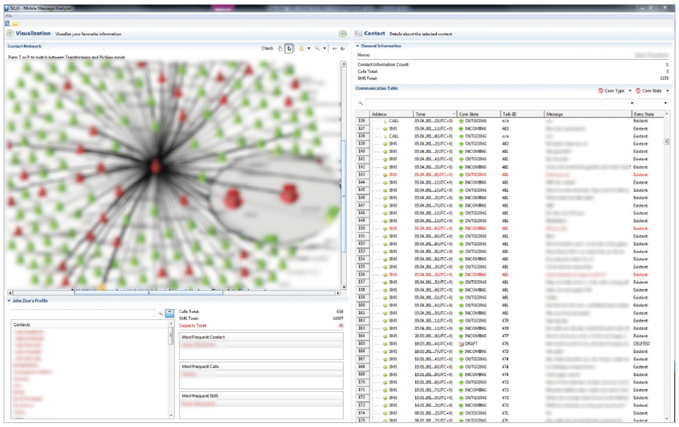


Figure 7: Dictionary containing pronunciation profiles as a basis for matching terms with high failure tolerance.

For testing the universality of the proposed process chain and especially the dictionary additional corpora are required. Fortunately, due to our cooperation with the local prosecutor’s office additional data is provided. Finally, initial development of an algorithm, which aims at calculating the conversation significance value with a high failure tolerance as shown in Fig. 7, is currently work in progress. Here, pronunciation profiles are used as a basis for understanding special terms.

VI. Performance of a Prototype

The implemented MoNA prototype show an F_1 score of about 0.80 for both string matching and phonetic matching algorithms in relevance classification of identified conversations. However, both algorithms show opposite performance with respect to sensitivity and recall (string matching: 1.0 sensitivity, 0.67 recall, phonetic algorithms: 0.67 sensitivity, 1.0 recall). In performance testing, a dictionary of keywords commonly used in the local drug scene of the western Saxony area had been provided by investigators with expert knowledge. As demonstrated in the Word Dictionary Potency section, coverage and potency of the provided dictionary is rather low, which is the cause for the discrepancy in recall, respectively, sensitivity and leaves room for improvement. Thus, future research and studies have to focus on keyword selection, dictionary development and refinement. Nevertheless, the workload for manual peruse and annotation has been reduced significantly to 15% by integrating the MoNA prototype into the investigation process chain.


M2A - Mobile Message Analyzer
File
Visualization
Visualize your favourite information
Contact Network
Check
Press T or P to switch between Transforming and Picking mode.
John Doe's Profile
Contacts
Contact
Details about the selected contact
General Information
Communication Table
Com Type
Com State

Calls Total: 610
SMS Total: 14307
Suspects Total: 60

Most Frequent Contact
[redacted]

Most Frequent Calls
[redacted]

Most Frequent SMS
[redacted]

Name: [redacted]
Contact Information Count: 1
Calls Total: 3
SMS Total: 2172

	Address	Time	Com State	Talk-ID	Message	Entry State
136	CALL	05.04.201..2(UTC+0)	OUTGOING	n/a	[redacted]	Existent
137	SMS	05.04.201..1(UTC+0)	INCOMING	483	[redacted]	Existent
138	CALL	05.04.201..8(UTC+0)	OUTGOING	n/a	[redacted]	Existent
139	SMS	05.04.201..2(UTC+0)	OUTGOING	482	[redacted]	Existent
140	SMS	05.04.201..2(UTC+0)	INCOMING	481	[redacted]	Existent
141	SMS	05.04.201..4(UTC+0)	OUTGOING	481	[redacted]	Existent
142	SMS	05.04.201..2(UTC+0)	INCOMING	481	[redacted]	Existent
143	SMS	05.04.201..9(UTC+0)	OUTGOING	481	[redacted]	Existent
144	SMS	05.04.201..1(UTC+0)	INCOMING	481	[redacted]	Existent
145	SMS	05.04.201..1(UTC+0)	OUTGOING	481	[redacted]	Existent
146	SMS	05.04.201..2(UTC+0)	INCOMING	481	[redacted]	Existent
147	SMS	05.04.201..7(UTC+0)	INCOMING	481	[redacted]	Existent
148	SMS	05.04.201..1(UTC+0)	OUTGOING	481	[redacted]	Existent
149	SMS	05.04.201..1(UTC+0)	OUTGOING	481	[redacted]	Existent
150	SMS	05.04.201..6(UTC+0)	INCOMING	481	[redacted]	Existent
151	SMS	05.04.201..6(UTC+0)	OUTGOING	481	[redacted]	Existent
152	SMS	05.04.201..8(UTC+0)	OUTGOING	481	[redacted]	Existent
153	SMS	05.04.201..8(UTC+0)	OUTGOING	481	[redacted]	Existent
154	SMS	05.04.201..8(UTC+0)	INCOMING	481	[redacted]	Existent
155	SMS	05.04.201..3(UTC+0)	OUTGOING	481	[redacted]	Existent
156	SMS	05.04.201..3(UTC+0)	INCOMING	481	[redacted]	Existent
157	SMS	05.04.201..5(UTC+0)	OUTGOING	481	[redacted]	Existent
158	SMS	05.04.201..0(UTC+0)	INCOMING	481	[redacted]	Existent
159	SMS	05.04.201..5(UTC+0)	OUTGOING	481	[redacted]	Existent
160	SMS	05.04.201..4(UTC+0)	OUTGOING	481	[redacted]	Existent
161	SMS	05.04.201..5(UTC+0)	INCOMING	481	[redacted]	Existent
162	SMS	05.04.201..1(UTC+0)	OUTGOING	479	[redacted]	Existent
163	SMS	05.04.201..1(UTC+0)	INCOMING	478	[redacted]	Existent
164	SMS	20.03.201..0(UTC+0)	INCOMING	477	[redacted]	Existent
165	SMS	19.03.201..4(UTC+0)	DRAFT	476	[redacted]	DELETED
166	SMS	18.03.201..7(UTC+0)	INCOMING	475	[redacted]	Existent
167	SMS	15.01.201..4(UTC+0)	OUTGOING	474	[redacted]	Existent
168	SMS	15.01.201..6(UTC+0)	OUTGOING	474	[redacted]	Existent
169	SMS	15.01.201..6(UTC+0)	INCOMING	474	[redacted]	Existent
170	SMS	15.01.201..6(UTC+0)	OUTGOING	473	[redacted]	Existent
171	SMS	15.01.201..7(UTC+0)	INCOMING	473	[redacted]	Existent
172	SMS	15.01.201..5(UTC+0)	OUTGOING	473	[redacted]	Existent
173	SMS	14.01.201..2(UTC+0)	INCOMING	472	[redacted]	Existent
174	SMS	08.01.201..2(UTC+0)	OUTGOING	471	[redacted]	Existent
175	SMS	08.01.201..2(UTC+0)	OUTGOING	471	[redacted]	Existent

Figure 8: MoNA user interface. The communication network is visualised and highlighted by colour in accordance to scored conversation relevance. Contact information and message histories are further reported and interactively explorable. Sensitive information regarding the closed criminal investigation is disguised.

VII. Conclusion and Future Work

Manual forensic peruse and analyses of SMS and IM messages is a time-demanding and error-prone process. In addition, in case of minor or moderate offences and crimes, such forensic investigations are not justifiable for economical reasons. In recent work it has been shown that automated strategies for information retrieval and mining in message corpora is difficult to realize due to information uncertainty and ambiguity introduced by grammatical and semantic structures usually uncommon in well-written and error-free texts. Existing computational text analyses approaches are predominantly tailored towards a clearly defined semantic domain and are employed to domain-specific corpora of semantically and grammatically correct texts. Successful utilisation of such techniques is thus often limited or even impossible in the context of forensic SMS and IM message analyses.

In this work, a computational approach is proposed that aims at reducing the amount of messages prior to manual peruse by identifying conversations in message histories, which might contain evidential information relevant in investigation. This approach initially identifies conversations in message histories based on statistical analyses of the characteristic behaviour of text communication between participants. Individual identified conversations are subsequently scored with respect to predicted crime-related relevance based on a key word dictionary deduced from practical knowledge of investigators. This evaluation is further used in conversation reporting and visualization within the communication network. As demonstrated, the implemented prototype, MoNA, shows acceptable performance in this respect. Although widely applied software (such as Oxygen Forensics [22], XRY Physical [23] and UFED Touch Ultimate [24]) provide valuable means for data extraction and visualization, the process of data exploration, annotation and peruse is still required to be conducted manually. Here, as a tool for case-based forensic semantic analyses, MoNA could provide a valuable missing link in the process chain. Furthermore, MoNA currently features a data interface to process results and data derived by means of UFED software packages. In the near future, here presented approaches are ought to be refined. Implementations of additional data interfaces compatible with software listed above are currently work in progress.

Acknowledgment

The authors would like to thank the members of the criminal investigation department and the prosecutorial Chemnitz (Germany). We acknowledge funding by the Free State of Saxony and the University of Applied Sciences Mittweida.

References

[1] M. Spranger, E. Zuchantke, and D. Labudde, “Semantic tools for forensics: Towards finding evidence in short messages,” in Proc. 4th. International Conference on Advances in Information Management and Mining, IARIA. ThinkMind Library, 2014, pp. 1–4.

[2] K. Barmpatsalou, D. Damopoulos, G. Kambourakis, and V. Katos, “A critical review of 7 years of mobile device forensics,” Digital Investigation, vol. 10, no. 4, 2013, pp. 323–349.

[3] A. Skudlark, “Characterizing SMS Spam in a Large Cellular Network via Mining Victim Spam Reports,” International Telecommunications Society (ITS) Biennial Conference, Tech. Rep., December 2014.

[4] I. Ahmed, D. Guan, and T. C. Chung, “Sms classification based on naïve bayes classifier and apriori algorithm frequent itemset,” International Journal of Machine Learning and Computing, vol. 4, no. 2, April 2014, pp. 183–187.

[5] Q. Xu, E. W. Xiang, Q. Yang, J. Du, and J. Zhong, “Sms spam detection using noncontent features,” IEEE Intelligent Systems, November/December 2012, pp. 44–51.

[6] D. G. A. Al-Talib and H. S. Hassan, “A study on analysis of sms classification using tf-idf weighting,” International Journal of Computer Networks and Communications Security, vol. 1, no. 5, October 2013, pp. 189–194.

[7] M. B. Deepshikha Patel, “Mobile sms classification: An application of text classification,” International Journal of Soft Computing and Engineering, vol. 1, no. 1, March 2011, pp. 47–49.

[8] S. Ishihara, “A forensic authorship classification in sms messages: A likelihood ratio based approach using n-gram,” in Proceedings of the Australasian Language Technology Association Workshop 2011, Canberra, Australia, December 2011, pp. 47–56. [Online]. Available: http://www.aclweb.org/anthology/U/U11/U11-1008

[9] T. Chen and M.-Y. Kan, “Creating a live, public short message service corpus: the nus sms corpus,” Language Resources and Evaluation, vol. 47, no. 2, 2013, pp. 299–335.

[10] D. H. W. Dannis Muhammad Mangan, “Information extraction from short text message in bahasa indonesia for electronics,” Jurnal Sarjana Institut Teknologi Bandung bidang Teknik Elektro dan Informatika, vol. 1, no. 1, April 2012, pp. 29–32.

[11] S. Cooper, R.L.and Manson, “Extracting temporal information from short messages,” in British National Conference on Databases, Glasgow, July 2007, LNCS 4587. LNCS, Springer, 2007, pp. 224–234.

[12] K. Nebhi, “Ontology-based information extraction from twitter,” in Proceedings of the Workshop on Information Extraction and Entity Analytics on Social Media Data. Mumbai, India: The COLING 2012 Organizing Committee, December 2012, pp. 17–22. [Online]. Available: http://www.aclweb.org/anthology/W12-5502

[13] VATM, “Number of SMS and MMS sent in Germany from 1999 to 2014* (in millions per day),” 2016, URL: http://www.statista.com/statistics/461700/number-of-sms-and-mms-sent-per-day-germany/ [accessed: 2016-01-03].

[14] G. Evans and V. Gosalia, “The coming storm: Companies must be prepared to deal with text messages on employee mobile devices,” Digital Discovery & e-Evidence, 2015.

[15] M. Spranger and D. Labudde, “Semantic tools for forensics: Approaches in forensic text analysis,” in Proc. 3rd. International Conference on Advances in Information Management and Mining (IMMM), IARIA. ThinkMind Library, 2013, pp. 97–100.

[16] 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.

[17] 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.

[18] H.-J. Postel, “Die Kölner Phonetik - Ein Verfahren zur Identifizierung von Personennamen auf der Grundlage der Gestaltanalyse.” IBM-Nachrichten, vol. 19, 1969, pp. 925–931.

[19] L. Philips, “The Double Metaphone Search Algorithm.” C/C++ Users Journal, vol. 18, no. 6, 2000, pp. 925–931.

[20] B. W. Matthews, “Comparison of the predicted and observed secondary structure of t4 phage lysozyme.” Biochim Biophys Acta, vol. 405, no. 2, Oct 1975, pp. 442–451.

[21] Cellebrite Mobile Synchronization LTD. UFED Physical Analyzer - Mobile Daten ermitteln, dekodieren und bereitstellen. [Online]. Available: http://www.cellebrite.com/Mobile-Forensics/Applications/ufed-physical-analyzer [accessed: 2016-03-01]

[22] Oxygen Forensics, Inc. Oxygen Forensics. [Online]. Available: http://www.oxygen-forensic.com/de/ [accessed: 2016-03-01]

[23] MSAB. XRY Physical. [Online]. Available: https://www.msab.com/products/xry/#physical [accessed: 2016-03-01]

[24] Cellebrite Mobile Synchronization LTD. UFED Touch - Eine hochleistungsfähige Lösung für hochleistungsfähige Geräte. [Online]. Available: http://www.cellebrite.com/de/Mobile-Forensics/Products/ufed-touch [accessed: 2016-03-01]