In Search for Intelligence: Automatically Estimating the Implicitness of Police Officers' Observation Messages, an Ongoing Action Research

Xandra van de Putte[a] Paul Oling[b] Jan-Kees Schakel[b]

[a] *Knowledge and expertise centre for intelligent data analysis (Kecida), Netherlands Forensic Institute, P.O. Box 24044, 2490 AA The Hague, the Netherlands*
[b] *National Police Services Agency (KLPD), P.O. Box 100, 3970 AC Driebergen, the Netherlands*

Abstract

This paper presents a method to identify observations containing information about potentially suspect situations (intelligence), made by police officers and registered in a digital notebook. The underlying assumption was that messages containing intelligence have a more implicit character than factual reports on car accidents and the like. To identify messages with a high level of implicit content, we applied a classification method using different kinds of features extracted from the text. Besides the representation of a document in word vectors, we also added additional syntactical and lexical features: number of adjectives, number of nouns, number of verbs and number of characters used in the message. One predefined field of the message, used for the time interval in which the observation took place, was also added. Moreover, domain knowledge, typical jargon used by police offers, was added to adjust the classifier such that the number of false negatives was minimized. The results show that the number of adjectives was of high value in the classifier and that adding domain knowledge to the classifier does help minimizing the number of observations falsely classified as having no implicit content.

1 Introduction

This paper presents part of an ongoing action research (AR) within the National Police Services Agency (KLPD). Among other tasks, the KLPD is responsible for the public safety and security of the national infrastructural networks (highways, waterways, railways and aviation). These traffic flows are characterized by high volumes, high velocity, and high dynamics, all contributing to anonymity. To identify criminal behavior or suspects within these flows the KLPD is presented with the challenge to organize for the real-time exchange and utilization of distributed information and knowledge. This is a two-sided challenge: officers working in the flow need to have access to information and expertise elsewhere in the organization, and individual observations of officers in the flow need to be accessible by the organization to identify possible relations. The aim of the ongoing AR and of this progress paper is to 'improve the observation capacity of the force and its subsequent ability to act selectively and in a timely manner' [1].

Organizing the KLPD in such a way that distributed information and knowledge can be exchanged successfully between geographically distributed, ad-hoc (virtual) teams such as between police officers in the field and experts in their offices, can be quite challenging. Based upon the ideas of collective memory [2], knowledge transfer can be enhanced through transactive memory systems (TMS) [3, 4]. TMS are shared systems people employ to divide responsibility, to keep each other informed and to access each other's knowledge [5]. TMS can be used to support the exchange of distributed knowledge. The concept of TMS acknowledges different types of knowledge, namely, codified knowledge, personalized

knowledge, and knowledge embedded in organizational structures, routines, methods, technology, etc. Personalized knowledge may be extremely hard to codify [6].

In 2007 the highway patrol started a pilot project to gather intelligence about suspect objects, subjects and situations, which the officers encountered during their work. The premises in this pilot project are that, 1) officers are capable of noting suspect situations based upon personalized knowledge, and 2) by codifying their observations in a digital notebook (rather than in an analogue one, as they used to do) the observations can be made accessible to the organization for further analysis. Not long after the introduction of the digital notebook the users started to log all their activities, effectively expanding its use. This behavior - known as bricolage [7] - may increase user value, but also increased the total amount of messages dramatically. This created a significant problem in the near real-time utilization of the originally sought after intelligence. Analyzing and selecting electronic messages manually is time consuming, while the value of the information reduces in time. Hence, an intelligent routine had to be developed to aid the swift selection of potentially interesting messages for further investigation by experts. In this paper we depart from the hypothesis that intelligence messages can be distinguished from routine activity reports by their relatively implicit nature, and hence, by being rather descriptive and lengthy – and that these characteristics should be detectable through text mining.

Using classifications, we try to identify the presence of implicit content within messages codified in the digital notebook. Within this paper we focus on these classification techniques and their added value to the near real-time analysis of potentially interesting messages.

In this paper we present a solution for automatically identifying messages that are relevant for immediate analysis, based upon their level of implicitness. The paper is organized as follows. Section two describes a set of hypotheses for the characterization of messages with implicit content. In section three we describe an appropriate feature set to perform the classification task and the method we used to build the classifier using the obtained features from a training set with manually classified messages. Based upon domain knowledge (policing), we minimized the false negatives and false positives made by the classifier by adding additional features. We also discuss empirical results. In the final section, we conclude and describe future work.

2 Characterization of Relevant Messages

The hypothesis behind our approach was that officers *circum*scribe rather than *des*cribe situations which potentially threaten safety or security, but which are not sufficiently clear to justify intervention. For the characterization of *circum*scriptions, we used the following hypotheses:

1) Messages describing suspect situations are considered to be longer than messages not describing suspect situations.
2) Certain terms are more likely to be used when describing suspect situations then when describing non-suspect situations. These terms also include domain specific words, or letter-combination, unknown to regular dictionaries but part of the policing jargon.
3) Certain punctuation marks indicate towards the officer's feelings about a situation. For instance, to accentuate something one might use one or more exclamation marks. A question mark can indicate if something is unclear or vague. Ellipses ('…') can be used when an officer has difficulties explicating his thoughts.
4) Suspicious situations are formulated in a more undetermined way. Hence, it may be expected that an officer would use particular syntactic classes differently, such as a higher amount of adjectives. This assumption is partly based upon the work of Wiebe [8] who used, among other features, the presence or absence of particular syntactic classes to separate subjective from objective sentences.

3 Method

In order to select and present the right selection of messages to analysts within the organization, we chose to classify each individual message to determine its relevance for the analyst. The dataset was gathered during the digital notebook pilot and some domain knowledge was gathered through a dozen observations in which one of the authors went along with officers from the traffic police during their surveillances, both obvious and discreet. By classifying each message using several features we expected to be able to grade each message on a scale ranging from 1 (positive, high potential to be relevant for further investigation) to 0 (negative, low potential to be relevant for further investigation).

The desired result was a classifier that could be interpreted and edited easily. The reason for this is that terms used for specific situations or objects may change over time. Even though the classifier will be updated regularly using new training data, this will not ensure immediate change of classification directly after the change of terms. For example, if the name of a specific police database changes, the officers could use the new name instead of the former name. To avoid missing important messages that contain such a feature, the feature must be added immediately to the classifier. There are two types of classifiers that can be used for this kind of task, namely rule and tree-based classifiers.

Section 3.1 describes the preparation of the data and the construction of the feature set from the data. An overview of this is given in figure 1. In section 3.2 we show how we built the classifier and refined it with domain knowledge. Section 3.3 discusses the results of the experiments.

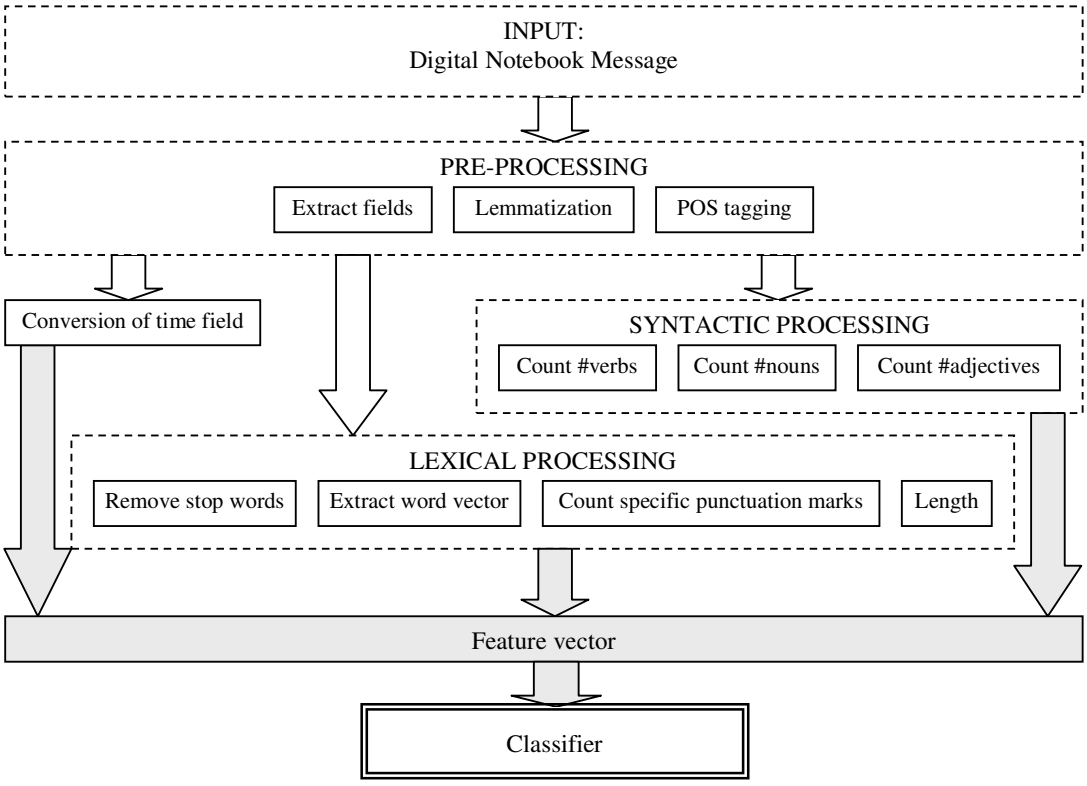

Figure 1: Overview of the extraction of features from a digital notebook message

3.1 Data Preparation and Feature Selection

The observations were registered through a mail system via Mobile Data Terminals (MDT) in police cars. Information had to be inserted in a web form with predefined fields. In one field the actual observation could be described, others were predefined for describing the type of observation, vehicles involved, officers' personal information, time of observation, etc. Almost 5000 messages were provided of which 200 were labeled as positive (high potential to be relevant for further investigation) and 207 as negative (low potential to be relevant for further investigation) on the scale as described in the previous chapter. The messages were collected in the period of December 2007 until April 2009. While the first version of the web form was intentionally unstructured and simple, in time additional fields were added to the form to simplify adding non-suspect, structured, situations. This development matches the bricolage process as previously described. Therefore, the first task in this experiment was to extract, combine and structure the data provided for further use.

The next step was to choose relevant fields for classification. Because the officers were free to fill out the information, not all fields were filled or filled properly. The following fields were chosen; 1) the actual observation, 2) type of observation and 3) time of observation. Because personal and vehicle information were considered irrelevant for this type of classification they were removed from the set. This information may become relevant after the classification when further investigation by experts is needed. The time of observation was converted into one of five intervals representing different timeslots throughout the day. These timeslots were carefully selected in a way that the two rush hours were in separate timeslots. During rush hour, it seemed that officers were more focused on other duties, namely the prevention of traffic jams and accidents. Therefore, the assumption was made that officers will be less likely to write down observations related to suspicious situations (security) during rush hours.

Furthermore, the observation and type of observation field had to be converted to word vectors. Stop words, place names and words containing both numbers and letters (such as license plates or names of motorways) were removed and the lemmas of the remaining words were used to build the vectors. The high dimensionality of the remaining data had to be reduced so that the desired type of classifier as described above (rule or tree based) could be build efficiently. The Odds Ratio was chosen to calculate the importance of a term [9]. For each appearing term, the Odds Ratio was calculated in both positive and negative perspective:

$$OddsRatio(t,x) = \frac{\sum d_x(t)(1 - \sum d_y(t)/D_y)}{\sum d_y(t)(1 - \sum d_x(t)/D_x)}$$

where t is the term t found in class n, x the class (positive or negative) for which the ratio is calculated, D_n the number of documents in class n and $d_n(t)$ a document in class n containing term t.

When the denominator is zero, the Odds Ratio cannot be calculated and the number of documents in which the term appeared was chosen. The top of the term lists, chosen using a threshold, of both positive and negative perspective was chosen as the feature set. This method addresses hypothesis 2 that certain terms are more likely to be used in positive rather than in negative messages and vice versa.

As discussed in section two, other properties of a description than the words being used can also be characteristic for the identification of implicit content. The use of relatively more words can indicate that it was not easy to express oneself (hypothesis 1). To describe something that is not obvious, more adjectives may be needed such as 'weird' and 'distinct' (hypothesis 4). Therefore, we also determined the length of each observation field and the number of adjectives, nouns and verbs that were used in the messages. Also, certain punctuation marks may indicate the implicitness of a message (hypothesis 3). The total number of exclamation marks, ellipses ('…') and question marks was also added as a feature. In total 176 features remained to build the base classifier. In table 1 the constitution of the feature types together with the methods used to obtain these features (automatically, semi-automatically) can be found.

3.2 Building the Classifier

As we described above, the desired classifier is a tree- or a rule-based classifier. We used Weka to build classifiers using four types of classification algorithms (M5P [10], REPTree [11], M5Rules [12] and

ConjunctiveRule [13]), which are commonly used tree and rule based algorithms. We then chose the one which best fitted the data using 10 fold cross-validation. M5P and M5Rules performed best (based upon the Root Mean Squared Error and the number of false negatives). Their performances were approximately the same, so we chose the one with the least complex result: M5Rules, see figure 2. The first rule of the result uses the number of adjectives. When the number of adjectives is higher than four, the following calculation adds 0.6226 (initial score) to the score and 16 features are used to adapt the score. Both characteristic words for the negative and positive set as the time of incident, number of adjectives, nouns and verbs and the length of the observation field, are used in the rule. The second rule, applied when there are less than 5 adjectives, uses a characteristic term of the negative set (because we used classified information, the terms used by the classifier cannot be given). If this term does not appear in the message, a similar calculation is done as the previous one, but with less features and a smaller initial score. The last rule results in a score of zero, meaning that the message is not likely to be potentially interesting for further investigation.

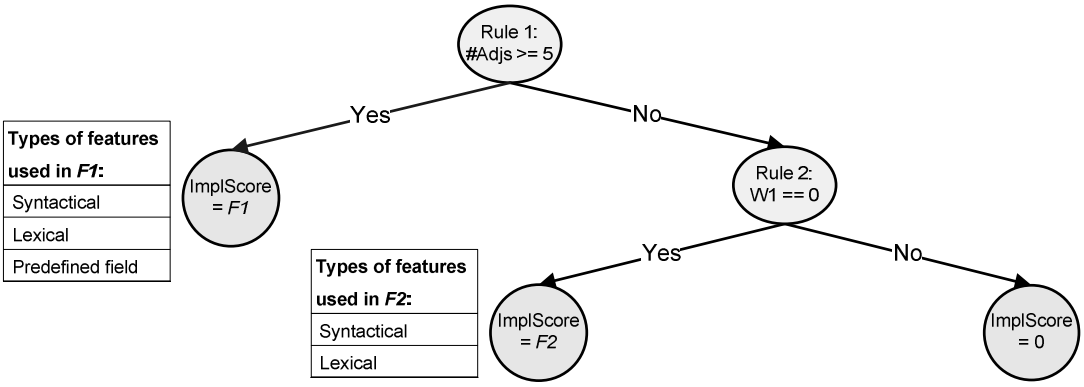

Figure 2: The output of the M5Rules classification algorithm. #Adjs and W1 mean the number of occurrences of adjectives and the number of occurrences of a specific term respectively. ImplScore means the likeliness that a message contains implicit information and F2 is the formula that contains specific features to calculate this likeliness.

Within the positive set 180 items received a score of 0.5 or higher. The negative set contained 26 items with a score higher than 0.5. Considering the goal of this research - to identify implicit content in messages - it is preferred to minimize the false negatives in order to preserve all messages that possibly contain some sort of implicit content. We studied the type of false negatives and denoted two types of messages. The most of these messages were short ones with a relatively small amount of adjectives. These messages contained different typical words for describing a suspicious object or situation. The next type exists of longer messages with relatively many nouns and verbs (which has a negative effect on the score), and containing a typical term used for traffic accidents (a non-suspicious event), but the core of the message was not an accident. These messages also contained typical suspiciousness words and some of these messages also contained the punctuation marks described in hypothesis 3. One way to avoid these types of false negatives is to apply cost-sensitive learning to each of the types separately (to avoid generalization) and combine the resulting classifiers. However, the numbers of messages in these types were too small to perform useful classification. Because the obtained scores of the false negatives were to scattered, thresholding could not solve this problem. Therefore, we chose to manually add features to the classifier that could minimize the false classifiers. After denoting the types of mistakes of the classifier, we used appropriate terms from the positive set which had a fairly high Odds Ratio. People not always use the same words for the same meanings. Based upon several field observations and interviews with police officers, we were able to use domain knowledge to select similar or related words of some features and included them in the appropriate feature. For example, the feature *odd* could change into the rule *max(odd, weird, suspicious)*, where the maximum number of occurrences of the term *odd*, *weird* and *suspicious* is used instead of the number of occurrences of the term *odd*. The added feature, which held

the number of particular punctuation marks used, did not occur in the automatically built classifier. Since a couple of false negatives contained this feature, we added this feature to another appropriate feature (using the max function). For this, we chose a low-weighted feature to avoid negative messages also containing these punctuation marks getting high scores.

Adding domain knowledge resulted in a lower number of false negatives, but the number of false positives increased. A large amount of false positives would result in a time-consuming task to select the more important messages. Therefore, we also did a similar (but smaller) addition to the features characteristic for the negative set used in the classifier. The final classifier contained 32 features, of which the numbers and selection method of the different types of features can be found in table 1.

Type of feature	**Selection method**	**#features in initial feature set**	**#features used in initial classifier**	**#features used in final classifier**
Nominal	Automatically	1	1	1
Regular vocabulary	Semi-automatically	170	11	11
Special vocabulary	Manually	0	0	15
Other lexical features (punctuation mark, etc.)	Automatically	2	1	2
Syntactic	Automatically	3	3	3
Total amount of features		**176**	**16**	**32**

Table 1: Types of features within the feature set with their selection method, constitution to the number of features used for building the initial classifier and constitution to the number of features used in the final classifier.

3.3 Empirical Results

The first hypothesis we made stated that positive messages are longer than negative messages. The built classifier used the length of the message in favor of the positivity of the message, but only a low weight was given to this feature. Adding domain knowledge, as addressed in the second hypothesis, aided minimizing the amount of false negatives. The third hypothesis stated that he number of occurrences of specific punctuation marks would occur more likely in a description of a suspect situations than in a description of a non-suspicious situation. Although it was not used by the initial classifier, the addition of it to the classifier helped reducing the number of false negatives. Finally, our fourth hypothesis was prominently present in the classifier. The number of adjectives was used in the first rule, which results a high initial score to be decreased or increased by presences of other features. Moreover, higher number of nouns and verbs resulted in a small subtraction of the score if the number of adjectives was higher than four, and higher number of adjectives resulted in a small addition. Hence, if more adjectives are used relative to the number of nouns and verbs, the score is more positive.

The final classifier was used to classify all observations provided. Of the 200 positive examples 7 items had a score lower than 0.5. These messages were manually reviewed and it was concluded that the implicit content of these messages was indeed low to absent. The number of false positives was 28. The false positives with the highest score contained the word ‘attention’, which was added by the officers to show the urgency of the message.

Of the whole set provided (4964 messages), 3612 items had a score lower than 0.5. This is in line with our expectation that most messages would not contain implicit content, partly because of the extensive bricolage that took place and the subsequent additional functions of the digital notebook. 815 Messages had a score higher than 0.7, which means they did contain a significant amount of implicit content.

4 Conclusion, Discussion and Future Work

The results show that using syntactical features, especially the number of adjectives, is very effective for identifying the presence of implicit information within officers’ observations. This makes it more robust for the change of word choices, for example when other officers are going to register their observations. Nevertheless it still needs to be updated and evaluated regularly to avoid missing positives. A drawback of using these types of syntactical features is that the technique used to determine the syntactical class of a word is sensitive to spelling, typing and grammatical errors. For example, when examining the false

positives with high scores, it turned out that words were wrongfully assigned as adjectives. Both unintended errors mentioned above, as wittingly choices for wrong spelling, namely the skipping of points in abbreviations, are mainly the cause of this. The latter could be solved by adding a vocabulary for abbreviations created and used by the officers involved, thus adding specific domain knowledge.

When further examining the false positives, some officers emphasized the urgency by using the word 'attention'. It seemed that the urgency could not be determined by the classifier, so the officer may have wrongly used the word 'attention'. The officer was not able to express the self-proclaimed urgency of his message in any other way. Also, some messages contained a high number of common adjectives that are not interesting for this classification (such as 'following'), which resulted in a high score. The latter did not occur often, but must be kept in mind for future implementation. Lists of common adjectives that are not interesting could be added to the stop word list for example. Furthermore, the usage of synonyms and more lists of relative terms used in the specific domain can be added to group the specific features. This way, different word choices for the same meaning result in the same contribution to the implicitness score.

To verify if the negatives are true negatives and the positives true positives, further evaluation has to be undertaken by experts, who may assign scores to a significant number of messages. This could eventually be done whilst using the classification in action.

To conclude a remark on the practical application of the above has to be made. By encoding knowledge, one should realise that the properties of knowledge will be changed such that it becomes easier to transfer [1]. When analysts are presented with these codified messages, they have to make sense of them. This starts with adding context. Hence, at this stage the attributes filtered out earlier, including e.g. the location and the name of the officer involved, become highly relevant. In the final solution these features need to be included.

It is expected that this research contributes to the enhanced utilization of distributed knowledge within the KLPD organization in general, and more particular within geographically distributed problem solving teams.

References

[1] Schakel, J.K. & Vries, E.J. de. The Organization of an Organizational Transactive Memory System for Nodal Governance to Fight Terror and Crime: an Action Research. *Submitted for publication.*

[2] Wegner, D. M. (1986). Transactive Memory: a Contemporary Analysis of the Group Mind. In: Mullen, B., & Goethals, G.R. (Eds.), *Theories of Group Behavior*, 185-208. Springer-Verlag, New York.

[3] Nevo, D. & Wand, Y. (2005) Organizational Memory Information Systems: a Transactive Memory Approach. *Decision Support Systems*, 39, 549-562.

[4] Oshri, I., Fenema, P. & Kotlarsky, J. (2008) Knowledge Transfer in Globally Distributed Teams: the Role of Transactive Memory. *Information Systems Journal*, 18, 593-616.

[5] Wegner, D.M., Raymond, & P., Erber, R. (1991) Transactive Memory in Close Relationships. *Journal of Personality and Social Psychology,* 61, 923-929.

[6] Orlikowski, W.J. (2002). Knowing in Practice: Enacting a Collective Capability in Distributed Organizing. *Organization Science,* 13, (3), pp. 249-273.

[7] Ciborra, C. (2002). *The Labyrinths of Information*. Oxford University Press Inc. New York.

[8] Wiebe, J., Bruce, R. & O'Hara, T. (1999). Development and Use of a Gold Standard Data Set for Subjectivity Classifications. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL-99)*, pp. 246-253.

[9] Mladenic, D. & Grobelnik, M. (1999) Feature Selection for Unbalanced Class Distribution and Naïve Bayes. *Proceedings of the Sixteenth International Conference on Machine Learning (ICML)*, pp. 258-267.

[10] Wang, Y. & Witten, I.H. (1997) Induction of Model Trees for Predicting Continuous Classes, *Proceedings of the European Conference on Machine Learning*, Praga, República Tcheca.

[11] Quinlan, J.R. (1993) *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, San Mateo, CA.

[12] Hall, M., Holmes, G., & Frank, E. (1999) Generating Rule Sets from Model Trees, *Proceedings of the 12th Australian Joint Conference on Artificial Intelligence*, Sydney, Australia, pp. 1-12.

[13] Kohavi, R. (1995) The Power of Decision Tables, *Proceedings of the European Conference on Machine Learning* 1995.