**Brief summary**

**Predictive text mining for cyber and digitized crime in police registrations**

**Goal**
In this study, a machine learning (ML) model was developed to classify police records in the *Basis Voorziening Handhaving* (BVH; the policy database that contains information on incidents reported to the police) that relate to cyber- or digitized crime. With that ML-model the extent of online crime in the BVH registration of 2016 is estimated. The background characteristics of linked suspects in these registrations of cyber and digitized crime have also been described. The research focused on registrations of three types of cybercrime (hacking, ransomware and DDoS attacks) and five types of digitized crime (online threats, stalking, online libel / slander / insult, online identity fraud and online buying and selling fraud).

**Method**
A ML-model has been developed that, on the basis of the texts in the BVH-2016 registration, is able to classify registrations in which one or more of the eight types of online crime are involved. The model has been trained and tested on a sample of registrations from the BVH. Subsequently, a large random sample of registrations from the BVH 2016 was used to estimate the total size of the registrations with the three cyber and the five digitized offenses and to describe characteristics of suspects involved in the registrations.

**Results**
- It is possible to develop a ML-model that accurately (precisely) classifies registrations concerning cyber and digitized crime in the BVH of the police for the year 2016.
- Based on a random sample from the BVH-2016 it is possible to estimate the registrations of eight types of online crime with a 95% confidence interval (i.e. quantification). In 2016 there were, with 95% confidence, between the 3,946-24,625 registrations relating to cybercrime and between 131,569-292,538 registrations relating to digitized crime. Because the data were based on a random sample and registrations of cybercrime were relatively rarer than registrations of digitized crime, the estimate of the first group was less accurate than that of the second group of offenses. These estimates were between the number of reports to the police of victimization of cybercrime and the number of crimes registered by the police.
- In order to describe background characteristics of suspects in online crime registrations, a very precise individual classification was required whereby all relevant registrations were also found (i.e. both high precision and recall). After all, it must be prevented that characteristics of suspects who are not related to online crime are counted and characteristics of suspects who do concern online crime are missed. Therefore, strict requirements must be set for this classification. The ML model was able to classify only registrations of hacking, ransomware and online purchase and sale fraud to a sufficient degree. The proportion of registrations in the BVH-2016 sample with these types of offenses was very low (less than 1% of the registrations).
The percentage of registrations of online offenses that had at least 1 suspect was lower for hacking and ransomware. For online purchase and sale fraud this percentage was higher than for the total number of registrations in the sample from the BVH-2016 (in 19%, respectively, 25 %, 60% and 42% of the registrations were at least one suspect). If there was a known suspect, the average number of suspects per registration for the online offenses is higher than for the total number of registrations of an offense with at least one suspect. For registrations concerning hacking, ransomware and online purchase / sales fraud with at least one suspect, the majority of the suspects are male and were born in the Netherlands. Of the registrations with at least one suspect the suspect was in 11%-33% a minor (hacking had the lowest and ransomware the highest percentage). However, these percentages were based very small numbers.

**Conclusion**
The results of this research showed that predictive textmining (PTM) is useful for accurately classifying registrations such as (one or more of the eight) online offenses. And under certain conditions, it is also possible to give estimates of the numbers of registrations concerning cyber and digitized offenses in the BVH-2016 within a 95% confidence interval. Because strict requirements have to be imposed on the precision of classifying to describe background characteristics of suspects, the model proved to be able to classify only registrations of hacking, ransomware and online purchase and sale fraud sufficiently accurately. We expect the developed ML model to be useful for trend research, but further research is needed. Due to expected changes in the appearance of cybercrime and changes in the (quality of the) registration source, it is necessary to update the model for other years.