MACHINE LEARNING METHODS FOR CLASSIFICATION OF SENSITIVE DATA

***Gints Rudusans**, **Gatis Vitols**
Latvia University of Life Sciences and Technologies, Latvia
*Corresponding author's email: gints.rudusans@gmail.com

Abstract
In the era of Big Data there are a lot of new challenges – understanding, processing, and securing the data, assuring data quality, dealing with data growth and other challenges. One of the challenges is to identify and classify data sets in different systems which must follow the conditions defined by different regulations. The classification of these data sets can be automated using machine learning methods. The aim of the research is to provide machine learning methods for classifying sensitive data. The research is based on analysis and comparison of European Union legislation and scientific literature, which addresses issues of data classification using machine learning methods. Special attention is paid to sensitive data defined by the General Data Protection Regulation (GDPR). The main focus in this research is on supervised learning algorithms, where one of the most effective is Naïve Bayes classifier. In order to achieve good results, there is a need to find a proper training data set. Usage of hybrid methods provides a new way for increasing performance of classifiers.
Key words: machine learning methods, classification of data, sensitive data.

Introduction

Data classification is essential to meet compliance standards and maintain control over sensitive data. Machine learning as a subset of artificial intelligence can help to identify and classify data; therefore, an important aspect is to understand the data regulations and rules that are applied for classifying the data.

Sensitive data is a set of special data categories, which must be processed with additional security. These special data categories include religious beliefs, political opinions, racial origin, ethnic origin, as well as membership of professional associations, data about individuals' genetics or biometry and data about sex life and sexual orientation.

Machine learning methods are used in many areas because they have the ability to solve complex issues and even make a prediction. Some of the researchers are working on identifying COVID-19 infection using images of chest X-ray (Sheykhivand *et al*., 2021). Face recognition (Wang *et al*., 2019) is also one of the tasks that machine learning can solve.

When it comes to classification of text, many of the researchers are trying to identify the hate speech in social networks. In one of such studies Wiedemann, Ruppert & Biemann (2019) developed an architecture of neural network for recognizing the hate speech. Some other related studies (Doostmohammadi, Sameti, & Saffar, 2019) involve subtasks: first task is to identify if the tweet is offensive or not, the other task is to determine whether the tweet is targeted and the last task is to determine to whom it is addressed.

The aim of the research is to provide machine learning methods for classifying defined by general data protection regulation.

Materials and Methods

This research contains analysis and comparison of European Union legislation, scientific literature related to sensitive data and machine learning for its classification. The types of machine learning are reviewed, and the Naïve Bayes classifier is explained on a simple example for identifying and classifying sensitive data.

Results and Discussions

The latest law in area of data protection is the GDPR – General Data Protection Regulation (European Parliament and Council, 2016), which within the European Union (EU) came into force on 25 May 2018. It is a huge step forward of achieving wider and more far-reaching protection of any individuals' personal data (Crutzen, Ygram Peters & Mondschein, 2019). The Regulation 2016/679 known also as GDPR, replaces the Data Protection Directive (DPD) 95/46/EC (European Parliament and Council, 1995) and Data Protection Act (DPA) 1998 (Parliament of the United Kingdom, 1998). It applies when personal data is being processed. The GDPR applies to any organization operating within the EU as well as any other organization outside the EU, but which offers services to customers or businesses in the EU. It means that almost any major corporation in the world is affected by GDPR and needs appropriate strategy. The Table 1 highlights some of the differences of definitions for private and sensitive data.

Companies that collect, use, or store personal data related to persons within the EU, must be compliant with the GDPR's privacy and security requirements or face large fines.

'Personal data' is a piece of information which relates to any person – either it's identified or identifiable. It is possible to identify the identifiable person in two ways – directly and indirectly, especially when using an identifier. In this case, as identifier can be used person's name, location data, an identification number, also one or many factors specific to the

Table 1

Difference between Data Protection Act (DPA) and General Data Protection Regulation (GDPR)

Type of data	DPA	GDPR
Personal data	Data related to a person who can be identified by data controller using its possessed data or other information	Any information from which it is possible to identify or potentially identify person.
Sensitive data	Personal data is considered sensitive in case of revealing: a) the racial or ethnic origin of the data subject, b) person's political opinions, c) beliefs of religion or beliefs like this, d) information about being a member of a trade union, e) conditions of person's physical and mental health, f) information about person's sexual life, g) the commission or possible commission by person of any offence, h) proceedings for the offense that is committed or possible offense	Data that have to be protected against authorized access, typically it consists of personal data revealing racial or ethnic origin, person's opinions about religion, politics, philosophical beliefs, or membership of professional associations, data about individual's genetics or biometry data, data containing information about persons health condition, sex life or sexual orientation
Processing of sensitive data	Special rules apply, but no special prohibition.	In addition to DPA conditions, the GDPR has few more, where at least one from all conditions must be satisfied: a) the condition about carrying out the obligations under employment has been expanded on the wording related to compliance with additions in obligations under collective agreement, social protection law or social security; b) the condition about establishing and exercising and defencing of legal claims was supplemented by wording in regards to data processing by courts that acts inside their juridical capacity c) the condition related to public interests inside the area of public health has been updated by providing a legal basis for regulatory uses of health data and by sharing health related data with providers of social care; d) new condition is added in addition to DPA which states that processing of sensitive data can also be performed for the purpose of archiving in the public interest, as well as the for research and statistics according to Article 89 (1).

genetic, mental, cultural, economic, physical, cultural or social identity of particular natural person.

The compliance with the GDPR means adopting internal processes to the requirements of regulation. It also means implementing data anonymization, minimization, classification, and other processes. To do a data classification, one should understand the data. However, the definitions of different data classification levels might differ between countries and districts.

The EU Data Protection Working Party during November 2014 published the main criteria that needs to be taken into account, when evaluating the requests for deletion, where the most important decision is about whether the personal data can be considered as sensitive (Li & Saxunova, 2020).

Figure 1 illustrates how the concepts of different types of personal data are related to each other. This classic relationship shows that some of the sensitive data overlaps with private data and its characteristics; therefore, it can be called private-sensitive data.

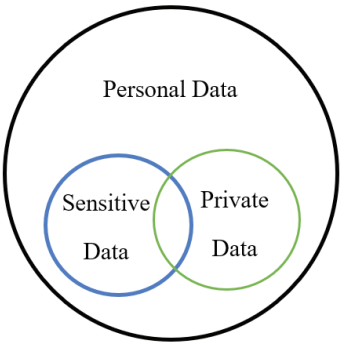

Figure 1. Classic relationship between Personal and Sensitive data.

Table 2

Data anonymization techniques
(Created by the authors based on Personal Data Protection Commission Singapore (2018) and Victor & Lopez (2020))

Techniques	Description
Attribute Suppression	The strongest technique of data anonymization, because an entire part of data is being deleted or a column is being removed from a dataset.
Record Suppression	Similar to attribute suppression, but more concrete – only specific records are deleted from a dataset.
Character Masking	Withing this technique the data is being hidden by altering its original values. Replacement is usually done by using symbols.
Pseudonymization	This technique makes data less accessible, but it does not fully remove the data and the individual still can be identified by using some additional or indirect information.
Generalization	This technique is known also as recoding – it removes a part of data, thus making the data values less precise.
Swapping	This technique is known as permutation or shuffling – the original data values are rearranged in order to keep the individual values inside the dataset but pointing them to non-original record.
Data Perturbation	The original values are slightly modified – rounding the numbers or by adding some random noise to original value.
Synthetic Data Generation	The synthetic (fake or artificial) data is generated by algorithms rather than by real events.
Data Aggregation	This technique instead of the original data is using its aggregated or average values.

After the GDPR was put into force, most of the companies sent a lot of e-mails to customers explaining the new private policy. The updated policy highlighted the rights to every single person to have the protection of its personal information and the rights to be forgotten, while the previous version of policy contained hundreds of legislative terminologies.

The process for identification sensitive data in most of cases is still done manually, which means higher cost and more time spent, as well as higher chance for getting errors. Moreover, inside the same organization but different departments, the same data can be classified in different ways, which makes it complicated to understand and process. Protecting user's sensitive data is a major challenge in many organizations. Recent data related incidents (McCandless, 2021) show that millions of users are affected by different data breaches and hacks. All kind of sectors are affected – academic, finance, gaming, government, military, health, etc. This clearly identifies the importance of data security. The first step in protecting sensitive data is identifying it. Once the data is classified, appropriate processes can be applied. Special routines for data anonymization – either in separate development environment or in general might be needed to implement. Data anonymization is a process where original data is either removed or replaced. Table 2 presents the techniques for data anonymization (Personal Data Protection Commission Singapore, 2018; Victor & Lopez, 2020). In order to automate the process of data identification and classification, machine learning methods can be used.

There are a lot of researches done on protecting sensitive data (Enck *et al.*, 2010; Rastogi, Chen, & Enck, 2013; Budianto *et al*., 2014). The researchers also have tried to solve different kind of data classification tasks. Some of the authors have tried to identify, weather the author of a written text is a male or female (Argamon *et al*., 2003). While many researchers try to solve more modern problems related to social network – identification of hate speech in Twitter and its main target (Ayo *et al*., 2020; Mulki *et al*., 2019).

The Figure 2 shows the types of machine learning (Betty Jane & Ganesh, 2019):

- supervised learning,
- unsupervised learning,
- reinforcement learning.

Supervised learning contains two variables – the input (x) and an output (Y). The algorithm is used for the function that maps an input to an output: $Y = f(x)$. The goal is to make a prediction for particular data by approximating the function of mapping until the output variable (Y) can be predicted by input variable (x).

It is called supervised learning because in the same manner as a teacher supervises the learning process, the algorithm is supervised by the training dataset, where input data already has a connection to right answers, in this case – the output data. Based on the answers that are provided, the algorithm keeps

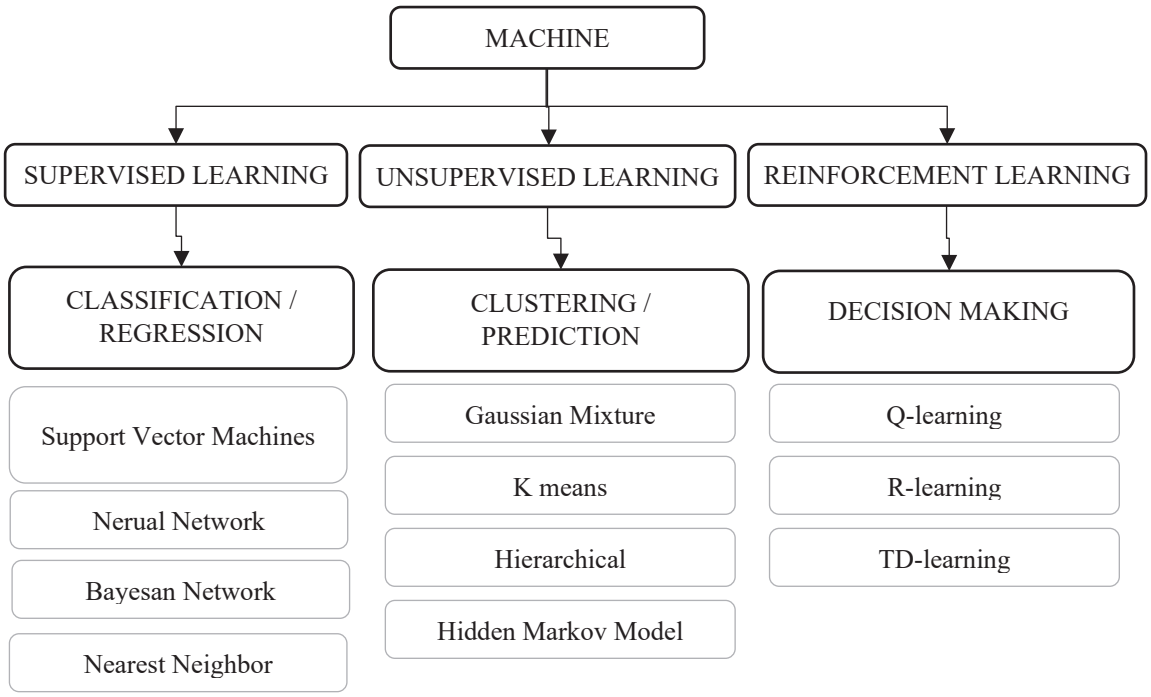

Figure 2. Machine Learning Techniques.
(Created by the authors based on Betty Jane & Ganesh (2019)).

Table 3

Example of training data

Text	Tag	Description
'He believes in God'	Sensitive	Sensitive information about religious belief
'John lives in Paris'	Non sensitive	Personal information about name and location
'Mary has got diabetes'	Sensitive	Personal information about name; Sensitive information that contains health data
Helen believes her soul will be in heaven	Sensitive	Personal information about name; Sensitive information about religious belief
'Her e-mail address is anna88@email.com'	Non sensitive	Personal information about email address;

training by predicting the output, that is corrected by the teacher. The learning process is completed once acceptable level of training results is achieved.

Classification and regression problems are the main subcategories of supervised learning. The problem of classification usually is when the output variable can be defined as a category, such as 'black' and 'white' or 'sensitive' and 'non sensitive'.

The Naïve Bayes classifier is popular and recognized among researchers (Barakaz, Boutkhoum. & Moutaouakkil, 2020). This algorithm is mostly used for solving problems of text classification and it is based on a probabilistic classification. It uses joint probabilities of categories and text division into words. Then, the algorithm makes an estimation of the final category where new data should be part of.

In the following example the Naïve Bayes classifier is explained on a simple example for identifying and classifying sensitive data. The data in Table 3 is the training data, where it is already classified in sensitive and non-sensitive data categories. When new data comes, for example 'Eric believes in reincarnation', we would need to classify it and tag if it is sensitive or non-sensitive. In this case, we see that this sentence represents information about person's name ('Eric') and his religious belief ('soul will be in heaven'). From this information we can understand that this information is sensitive, so now we should make sure, that the Naïve Bayes classifier understands the same. Since the Naïve Bayes classifier is a probabilistic classifier, then a probability of sentence 'Eric believes in reincarnation' should be calculated – first assuming that

it contains sensitive data and then that it is non-sensitive information. The higher probability should point to the correct classification. Written mathematically, we want to calculate P (sensitive | Eric believes in reincarnation) – the probability that the tag of this sentence is Sensitive given that the sentence is 'Eric believes in reincarnation'. To start building machine learning model, first, it is needed to decide on the features, which are information pieces that are taken from the text and given to the algorithm. For example, in case of sensitive data if we look at health-related data, the features could be person's gender, age, weight, etc. some of the known information could be also excluded, since it isn't useful, like person's favorite color. In a particular case, there are no numeric features, but only the text, which can't be directly used in calculations. Therefore, word frequencies are used – the order of words and sentence construction is ignored. The features in this case are the counts of each of the words.

Mathematically, the Naïve Bayes' theorem can be expressed as follows:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

In the example we have P (sensitive | Eric believes in reincarnation). Now, using the theorem we get:

$$P(sensitive|ericbelievesinreincarnation) =$$
$$= \frac{P(ericbelievesinreincarnation|sensitive) \times P(sensitive)}{P(ericbelievesinreincarnation)}$$

In this case the divisor can be removed, since for the classifier we need to find out the tag with biggest probability. Therefore, we can just compare

$$P(ericbelievesinreincarnation|sensitive) \times P(sensitive)$$

with

$$P(ericbelievesinreincarnation|nonsensitive) \times$$
$$\times P(nonsensitive)$$

The problem is that the sentence 'Eric believes in reincarnation' doesn't appear in our training data, so we get 0 for the probability both – in sensitive and non-sensitive calculations. In the theory of Naïve Bayes, we assume that each word in the sentence is independent, and we look at each word individually rather than looking at the whole sentence. In other words, our sentence 'Eric believes in reincarnation' has the same meaning and sensitivity as 'in reincarnation believes Eric' or 'in Eric reincarnation believes'. Hence, we can write this as:

$$P(ericbelievesinreincarnation) = P(eric) \times$$
$$\times P(believes) \times P(in) \times P(reincarnation)$$

And apply this to what we had before:

$$P(ericbelievesinreincarnation|sensitive) =$$
$$= P(eric|sensitive) \times P(believes|sensitive) \times$$
$$\times P(in|sensitive) \times P(reincarnation|sensitive)$$

One problem that appears at this step is that not all words appear in our training data. That leads to incorrect calculations for multiplication of probabilities – if probability of non-existing word is 0, then the multiplication of all probabilities will be 0 as well. To avoid this, we will add +1 to each of every word count from the training data. Hence, if we get a new word, that does not exist in our training data, we will assume, that the count of it is 1. The next steps are just calculations of probabilities. First, we calculate probability of each tag. For a given sentence in the training data, the probability that it is Sensitive P(Sensitive) is 3/5. Next is probability that it is Non sensitive P(Non sensitive), and it is 2/5. Further, we do a calculation based on the counts – in total, there is 21 unique word in our training data (24 non-unique). In order to calculate the probability for 'believes' we get:

$$P(believes|sensitive) = \frac{2+1}{15+21}$$

Table 4 shows the probabilities for all words met in sentence 'Eric believes in reincarnation' assuming they are sensitive or either non-sensitive.

Table 4

Probabilities for all words met in sentence 'Eric believes in reincarnation'

Word	P (word | sensitive)	P (word | non sensitive)
Eric	$\frac{0+1}{15+21} = 0.027777778$	$\frac{0+1}{9+21} = 0.033333333$
believes	$\frac{2+1}{15+21} = 0.083333333$	$\frac{0+1}{9+21} = 0.033333333$
in	$\frac{2+1}{15+21} = 0.083333333$	$\frac{1+1}{9+21} = 0.066666667$
reincarnation	$\frac{0+1}{15+21} = 0.027777778$	$\frac{0+1}{9+21} = 0.033333333$

The last step is to multiply all the probabilities met in each tag:

$$P(sensitive) \times P(eric|sensitive) \times \\ \times P(believes|sensitive) \times P(in|sensitive) \times \\ \times P(reincarnation|sensitive) = 0.000003215$$

$$P(nonsensitive) \times P(eric|nonsensitive) \times \\ \times P(believes|nonsensitive) \times P(in|nonsensitive) \times \\ \times P(reincarnation|nonsensitive) = 0.000000988$$

The results show that sentence 'Eric believes in reincarnation' is classified as sensitive, because it has higher probability, which is as expected since the test sentence contains information about religious belief.

Conclusions

1. In this research the main focus is on supervised machine learning algorithms. One of the simplest and most effective is Naïve Bayes classifier. It is fast and produces good results. However, when we talk about sensitive data and data regulations, we might need to look at some other algorithms or a combination of algorithms.
2. An important aspect in machine learning processes is the training data. Some methods might show better results when are trained using one training data set, but the result might become worse, if using other training data. Getting a good training data set for sensitive data is also a challenging task.
3. In order to achieve even better results, the machine learning methods can be used together with other methods – then they become hybrid methods. This concept, where classifiers are combined, provides a new way for increasing performance of the classifiers.

References

Argamon, S., Koppel, M., Fine, J., & Shimoni, A.R. (2003). Gender, Genre, and Writing Style in Formal Written Texts. *Text & Talk*. 23(3), 321–346, DOI: 10.1515/text.2003.014.

Ayo, F.E., Folorunso, O., Ibharalu, F.T., & Osinuga, I.A. (2020). Machine learning techniques for hate speech classification of twitter data: State-of-the-art, future challenges and research directions. *Computer Science Review*. 38, 100311. DOI: 10.1016/j.cosrev.2020.100311.

Barakaz, F.E., Boutkhoum, O., & Moutaouakkil, A.E. (2020). A hybrid naïve Bayes based on similarity measure to optimize the mixed-data classification. *TELKOMNIKA Telecommunication, Computing, Electronics and Control*. 19(1), 155–162. DOI: 10.12928/TELKOMNIKA.v19i1.18024.

Betty Jane, J., & Ganesh, E.N. (2019). A Review On Big Data With Machine Learning And Fuzzy Logic For Better Decision Making. *International Journal of Scientific & Technology Research*. 8(10). 1121–1125.

Budianto, E., Jia, Y., Dong, X., Saxena, P., & Liang, Z. (2014). You can't be me: Enabling trusted paths and user sub-origins in web browsers. In: Stavrou A., Bos H., Portokalidis G. (eds). Research in Attacks, Intrusions and Defenses. September 2014 (pp. 150–171). Springer International Publishing, Cham. DOI: 10.1007/978-3-319-11379-1_8.

Crutzen, R., Ygram Peters, G.J., & Mondschein, C. (2019). Why and how we should care about the General Data Protection Regulation. *Psychology & Health*, 34(11), 1347–1357. DOI: 10.1080/08870446.2019.1606222.

Doostmohammadi, E., Sameti, H., & Saffar, A. (2019). Ghmerti at SemEval-2019 Task 6: A Deep Word-and Character-based Approach to Offensive Language Identification. In: Proceedings of the 13th International Workshop on Semantic Evaluation, June 2019 (pp. 617–621). Minneapolis, Minnesota, USA: Association for Computational Linguistics. DOI: 10.18653/v1/S19-2110.

Enck, W., Gilbert, P., Chun, B.G., Cox, L.P., Jung, J., McDaniel, P., & Sheth, A.N. (2010). Taintdroid: An information-flow tracking system for realtime privacy monitoring on smartphones. In Proceedings of the 9th USENIX Conference on Operating Systems Design and Implementation. June 2014 (pp. 393–407). OSDI'10, USENIX Association, Berkeley. DOI: 10.1145/2619091.

European Parliament and Council (1995). Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data. Retrieved March 3, 2021, from https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:31995L0046.

European Parliament and Council (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC. Retrieved March 3, 2021, from https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:02016R0679-20160504.

Li, Y., & Saxunova, D. (2020). A perspective on categorizing Personal and Sensitive Data and the analysis of practical protection regulations. *Procedia Computer Science*, 170, 1110–1115, DOI: 10.1016/j.procs.2020.03.060.

McCandless, D. (2021). *World's Biggest Data Breaches & Hacks*. Retrieved February 25, 2021, from https://www.informationisbeautiful.net/visualizations/worlds-biggest-data-breaches-hacks/.

Mulki, H., Haddad, H., Ali, C.B., & Alshabani, H. (2019). L-HSAB: A levantine twitter dataset for hate speech and abusive language. In Proceedings of the Third Workshop on Abusive Language Online, July 28–August 2, 2019 (pp. 111–118). Florence, Italy. DOI: 10.18653/v1/W19-3512.

Parliament of the United Kingdom (1998). Data Protection Act 1998. Retrieved March 3, 2021, from https://www.legislation.gov.uk/ukpga/1998/29.

Personal Data Protection Commission Singapore. (2018). *Guide To Basic Data Anonymisation Techniques*. Retrieved December 12, 2021, from https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Other-Guides/Guide-to-Anonymisation_v1-(250118).pdf?la=en.

Rastogi, V., Chen, Y., & Enck, W. (2013). Appsplayground: Automatic security analysis of smartphone applications. In: Proceedings of the Third ACM Conference on Data and Application Security and Privacy. February 2013 (pp. 209–220). CODASPY '13, ACM, New York. DOI: 10.1145/2435349.2435379.

Sheykhivand, S., Mousavi, Z., Mojtahedi, S., Yousefi Rezaii, T., Farzamnia, A., Meshgini, S., & Saad, I. (2021). Developing an efficient deep neural network for automatic detection of COVID-19 using chest X-ray images. *Alexandria Engineering Journal*, 60(3), 2885–2903. DOI: 10.1016/j.aej.2021.01.011.

Victor, N., & Lopez, D. (2020). Privacy Preserving Sensitive Data Publishing using (k,n,m) Anonymity Approach. *Journal of Communications Software and Systems*. 16(1). 45–56. DOI: 10.24138/jcomss.v16i1.825.

Wang, P., Su, F., Zhao, Z., Guo, Y., Zhao, Y., & Zhuang, B. (2019). Deep class-skewed learning for face recognition. *Neurocomputing*. 363, 35–45. DOI: 0.1016/j.neucom.2019.04.085.

Wiedemann, G., Ruppert, E., & Biemann, C. (2019). UHH-LT at SemEval-2019 Task 6: Supervised vs. Unsupervised transfer learning for offensive language detection. In: Proceedings of the 13th International Workshop on Semantic Evaluation, June 2019 (pp. 782–787). Minneapolis, Minnesota, USA: Association for Computational Linguistics. DOI: 10.18653/v1/S19-2137.