ELRC Action: Covering Confidentiality, Correctness and Cross-linguality

**Tom Vanallemeersch,[1] Arne Defauw,[1] Sara Szoc,[1] Alina Kramchaninova,[1]
Joachim Van den Bogaert,[1] Andrea Lösch[2]**
[1]CrossLang
Kerkstraat 106, 9050 Gent, Belgium
{firstname.lastname}@crosslang.com
[2]DFKI
Stuhlsatzenhausweg 3, D-66123 Saarbrücken, Germany
andrea.loesch@dfki.de

Abstract

We describe the language technology (LT) assessments carried out in the ELRC action (European Language Resource Coordination) of the European Commission, which aims towards minimising language barriers across the EU. We zoom in on the two most extensive assessments. These *LT specifications* do not only involve experiments with tools and techniques but also an extensive consultation round with stakeholders from public organisations, academia and industry, in order to gather insights into scenarios and best practices. The LT specifications concern (1) the field of automated anonymisation, which is motivated by the need of public and other organisations to be able to store and share data, and (2) the field of multilingual fake news processing, which is motivated by the increasingly pressing problem of disinformation and the limited language coverage of systems for automatically detecting misleading articles. For each specification, we set up a corresponding proof-of-concept software to demonstrate the opportunities and challenges involved in the field.

Keywords: anonymisation, fake news, machine translation

1. Introduction

The development of language technologies (LTs) requires substantial amounts of language data. European, national and regional public services in all EU Member States continuously deal with a huge amount of multilingual textual information in original and translated form. With the European Language Resource Coordination (ELRC)[1] action, the European Commission (EC) has taken a decisive step towards minimising language barriers across the EU, through various types of activities: data collection, setup of a helpdesk for data providers, awareness raising, and assessing tools and techniques in specific LT areas. Within ELRC, the assessment of tools and techniques in these areas involves documenting them in a hands-on manner and performing corresponding tests and experiments, in order to inform EC staff and EU Member State representatives about the feasibility and impact of such technologies. The overall goal is to create proof-of-concept environments integrating relevant tools and services in order to facilitate their uptake by users from the public sector.

This paper aims to illustrate the development of two major LT assessments in ELRC, i.e. the Automated Anonymisation specification and the Multilingual Fake News Processing specification. The paper is structured as follows. In Section 2, we provide details on the ELRC action, the activities performed in it, and the methodology employed as part of the development of the LT specifications. In Sections 3 and 4, we zoom in on the aforementioned two specifications: (partially) automated anonymisation of monolingual or bilingual information and fake news processing with a multilingual focus. Finally, we provide conclusions in Section 5.

2. Overview of ELRC

The ELRC action was set up through the SMART 2014/1074 programme of Connecting Europe Facility (CEF) in April 2015 and is coordinated since then by DFKI,[2] in partnership with ELDA,[3] ILSP/Athena RC,[4] the Latvian company Tilde,[5] and the Belgian company CrossLang.[6] It is governed by the Language Resource Board (LRB), which consists of leading technological and public service representatives for each CEF affiliated country.

The data collection activities in the action (Smal et al., 2020; Lösch et al., 2021) include the maintenance of a language data sharing facility, the ELRC-SHARE repository.[7] They further include the setup of a Technical and Legal Helpdesk, which offers legal advice to data providers and turns their data into standardised, machine-readable formats and actionable language resources (LRs). The ELRC consortium also collects re-

[1] http://www.lr-coordination.eu

[2] Deutsches Forschungszentrum für Künstliche Intelligenz, Germany (http://dfki.de/en)

[3] Evaluations and Language Resources Distribution Agency, France (http://elda.org/en)

[4] Institute for Language and Speech Processing/Athena Research Centre, Greece (http://www.ilsp.gr/en)

[5] http://tilde.com

[6] http://www.crosslang.com

[7] https://elrc-share.eu

sources itself, more specifically parallel corpora involving a variety of languages, and makes them available through the repository. Through its data collection efforts, the ELRC action enables the development of European LTs for languages including (but not limited to) all EU official languages and directly contributes to improving the quality, coverage and performance of CEF eTranslation and other MT systems that need multilingual LRs as training data.

As for awareness-raising activities, they take the form of events (country-specific workshops, EU-wide technical workshops, and European conferences) and social media campaigns. They promote the importance of data sharing, spread information on the opportunities and challenges of specific types of LTs, and enable discussions among researchers, developers, and potential users.

Activities involving the assessment of tools and techniques relate to the following LT areas:[8] computer-aided translation, MT-supported translation of websites, automated classification of documents using a fixed taxonomy of labels, anonymisation of monolingual documents or parallel data, and detection of fake news articles in multiple languages. As part of these activities, various tools and techniques are tested and documented in a hands-on way, experiments are performed through various adaptations and configurations, and proof-of-concept environments are set up integrating tools, for instance in the form of dockers including software and deep learning models, in order to allow EC staff, EU Member State representatives, or other potential users to test tools and various configurations in a user-friendly way, and, if desired, to train new models. In case of anonymisation and detection of fake news articles, the consortium also organised consultation rounds involving academia, industry, and public organisations, in order to gather insights on scenarios, standards, best practices, and current challenges. These rounds allow for a more comprehensive view on the LT areas concerned and for reflecting this view in a proof-of-concept environment.

In the subsequent sections, we discuss the assessments involving a consultation round: the specification on anonymisation, which in its final stages, and the one on multilingual fake news processing, which is ongoing work.

3. Anonymisation

In this section, we motivate ELRC's anonymisation specification, describe tools and techniques, discuss the consultation round and its findings, identify scenarios for anonymisation, and provide details on the proof-of-concept software.

3.1. Motivation

The process of anonymisation consists of removing personal identifiable information (PII). It involves removing elements that can be used to identify a person, e.g. names, account numbers, mortgage amounts, etc., in a way that the resulting data cannot be associated with any individual. This process is also referred to as deidentification of data. It is important to note that some elements may be direct identifiers (such as a person's name) while others may only be indirect identifiers. For instance, more than 80% of the US population are likely to be uniquely identified based on a combination of three indirect identifiers, i.e. 5-digit ZIP code, gender, and date of birth (Sweeney, 2000).

While anonymisation focuses on the removal of PII, pseudonymisation involves the replacement of a named entity (NE) or pattern (e.g. numerical) by another element (a pseudonym). From a technical point of view (when automating processes), the term anonymisation carries the same meaning as pseudonymisation, but they have a different meaning from a legal point of view (European Union Agency for Cybersecurity, 2019): in contrast to pseudonymisation, anonymisation involves the *irreversible* altering of personal data in such a way that a data subject can no longer be identified directly or indirectly.

Anonymisation is particularly important when it comes to the sharing of language data. For instance, in case an organisation wants to store data and share it with other organisations without violating the General Data Protection Regulation (GDPR)[9] or in case data shall be processed and used for training a machine translation (MT) system. The GDPR lays out principles for the sharing and re-use of data that contains personal information, such as "purpose limitation" (the data should be collected for a specific purpose), "data minimisation" (only the data necessary for that purpose should be collected and processed) and "storage limitation" (the data should be stored for no longer than necessary for that purpose). When archiving information in the public interest, for scientific or historical research purposes, or for statistical purposes, measures like the use of pseudonyms need to be taken in order to respect the principle of data minimisation.

3.2. Tools and techniques

In the context of the ELRC action, we investigated the anonymisation of unstructured textual content (consisting of *running text*, i.e. sentences or paragraphs), rather than structured textual content in databases or spreadsheets, in which the fields and cells typically identify the type of information unambiguously. Looking at unstructured content, anonymisation involves two steps, (1) detecting what should be anonymised and (2) determining how it should be anonymised. The first step comprises the identification of person names, locations,

[8] Additional areas will be investigated in the course of 2022.

[9] `http://data.europa.eu/eli/reg/2016/679/2016-05-04`

dates, etc. The second step focuses on strategies like blackening NEs (e.g. by replacing a person name or location with a generic placeholder *X* or with an encrypted string), replacing NEs with a label indicating the type of information (e.g. *PERSON* or *LOCATION*, or a label inside a hierarchy like *NAME* → *given name*), or replacing NEs with a similar word (e.g. replacing a person's name with another person name). Table 1 shows an example of a text anonymised using these three different strategies.

Original sentence: **Jamie's manager Alice was disappointed when we has fired by TextCorp last Monday. Three days later, Jamie submitted a complaint to TextCorp.**
Option 1: X manager X was disappointed when he was fired by X last X. Three days later, X submitted a complaint to X.
Option 2: PERSON_1's manager PERSON_2 was disappointed when he was fired by ORGANISATION_1 last DAY_1. Three days later, PERSON_1 submitted a complaint to ORGANISATION_1.
Option 3: Peter's manager Roberta was disappointed when he was fired by IBM last Thursday. Three days later, Peter submitted a complaint to IBM.

Table 1: Three replacement strategies for anonymising

While manual anonymisation might be an option for small amounts of data, it becomes infeasible when dealing with large volumes. In that case, human intervention should be minimised. This is especially important in the multilingual context of the EU institutions (as well as that of European LT providers), which produce enormous volumes of language data covering many areas of life and different types of data. As such, when automating anonymisation with the goal of making the data re-usable e.g. for MT training, the first of the two steps mentioned earlier involves running a named-entity recognition (NER) system. The second step involves applying some replacement strategy to all NEs or a subset of them.

NER systems make use of machine learning models that are trained using sentences with manually annotated NEs, lists of known NEs (*gazetteers*), lists of patterns covering NEs (e.g. patterns for dates or email addresses), or a combination of these resources. State-of-the-art NER models rely on deep learning. While NER results are never perfect, state-of-the-art systems sometimes provide high-quality results; for instance, an F1 score of 93.79[10] was reported for a general-purpose NER model on an often used English gold standard (Stanislawek et al., 2019). NER models may be trained on domain-specific material, for instance on clinical reports (Abadeer, 2020).

As a replacement strategy, all NEs may be substituted by pseudonyms of some type (e.g. NE labels), or substitution may be limited to a subset, by applying a differential privacy technique. This type of technique has a mathematical foundation (Dwork, 2006). It adds noise to data in order to reduce the risk of reidentifying a specific individual, while keeping the data useful to a certain extent for further processing (e.g. to calculate statistics). When applied to unstructured text, it can be used to replace some NEs with similar ones and to keep others unmodified. A probability value can be set, controlling the number of replacements and hence the trade-off between privacy and usability (a probability of 1 leads to replacement of all NEs, and therefore to the highest privacy).

There are a few publicly available anonymisation tools for unstructured texts, in particular software resulting from EU-funded projects, such as TM-Anonymizer, developed in the CEF Data Marketplace project (Kamran et al., 2020), Text Transformer (Adelani et al., 2020), developed in the H2020 project COMPRISE, and Biroamer, which is part of the Bitextor package (Bañón et al., 2020) developed in the CEF Paracrawl project. These three tools have been integrated into the proof-of-concept software described in Section 3.5.

3.3. Consultation round

The consultation round consisted of organising meetings with central stakeholders concerned with automated anonymisation, in order to collect requirements and feedback from a variety of sources and application areas:

- Consortia of projects related to NER and anonymisation, such as MAPA,[11] ELG,[12] and COMPRISE.[13]

- Translation technology experts: the eTranslation developer team at Directorate-General (DG) Translation of the EC, as well as a representative from industry.

- Domain experts, in the field of anonymising legislative and legal documents (University of Bologna), police reports (University College London), and health-related texts (Vicomtech, a Spanish technological centre specialising in AI, visual computing, and interaction).

[10] Score calculated for the whole set of predictions, i.e. predictions with label *person*, *organisation*, etc.

[11] `https://mapa-project.eu`: a CEF Generic Services project in which a multilingual anonymisation toolkit for public administrations was constructed.

[12] `https://european-language-grid.eu`: European Language Grid, a platform for LTs in Europe.

[13] `https://www.compriseh2020.eu`: *Cost-effective, Multilingual, Privacy-driven voice-enabled Services*, a project working on the anonymisation of speech data and of transcribed speech.

- Digital Service Infrastructures of the EC, i.e. teams that deploy trans-European digital services in various domains and show an interest in the anonymisation of their data.
- The LRB of ELRC (see Section 2).

One of the main findings of the consultation round consists of the fact that the sensitivity of the data to be anonymised has an influence on the choice of replacement strategy. In case of high sensitive data, replacing NEs by other NEs may be preferred over replacement using NE labels in order to make it harder for a potential attacker to identify what text parts are the result of replacement.[14] The choice of the replacement strategy equally depends on downstream use. In case the anonymised text will be fed to another application, such as a summariser or a tool deriving some type of statistics, the use of *X* as a generic label or the use of named NE labels may be suitable. In case the data is targeted towards a human reader, substituting NEs by other NEs may be more useful. However, this is more complex: NEs have to be inflected in some languages and the coherence of the text should be ensured. For instance, a misfit between pseudonym and pronoun as in the following sentence should be avoided:

> *John won the game and her supporters cheered.*

Related findings concern the influence of the data sensitivity on the feasibility of automating anonymisation and on the type of evaluation. The application of an NER system to high sensitive data may not be sufficient to obtain a satisfactory result. For instance, in a police report where person names have been anonymised, certain events in the non-anonymised part of the text may still allow for revealing the person reported about. This may lead towards the need for an extensive manual review of the results of automated anonymisation. As for evaluation, an environment with low sensitive data may merely need a manual evaluation of sentences with automatic NE annotations in order to estimate the quality of the system, whereas a usage-based evaluation may be necessary in environments with high sensitive data, e.g. the organisation of a task in which a person without access to specific databases tries to track down the person being reported on. This type of evaluation allows assessing the deanonymisation (i.e. reidentification) risk.

From the consultation round, it is also apparent that most anonymisation tools and techniques involve monolingual data. Some tools allow for anonymising translation memories, by anonymising source and target sentences separately or by taking into account translation-equivalent words in source and target sentences (see the Biroamer tool discussed in Section 3.5). Concerning the construction of MT systems that are fit to translate anonymised sentences, a lot of exploration is still needed. For instance, one could replace NEs with their NE labels in the MT training data and in the MT input, but, in case of highly inflected languages, information on the NE's function should be taken into account in order to inform the MT construction process. After applying MT to the anonymised data, the organisation controlling the original data may want to deanonymise the MT output for internal use, based on the mapping table that links original text parts and pseudonyms. This deanonymisation involves copying the NE or translating it (e.g. in case of an organisation name), and possibly choosing an inflection.

The final major finding of the consultation round is that the user of an anonymisation system needs sufficient control on the deidentification process. The process ideally takes place at the user's premises because of confidentiality (which has implications on infrastructure, as deep learning models are involved) and the user needs to be able to specify a list with NEs (gazetteer) and patterns for elements to be replaced. Moreover, the user needs to have the possibility to manually correct the anonymisation results in an editor.

3.4. Scenarios

Based on the above findings, a number of basic and advanced scenarios in the area of anonymisation were identified.

One basic scenario, illustrated in Figure 1, involves a workflow in which a document is processed using NE labels. The document (e.g. a docx file) is preprocessed by separating text and layout, the text is segmented into sentences, and the latter are anonymised by running a deep learning model (neural network, NN) for NER. The workflow also consults a list of NEs and regular expressions specific to the organisation in question. The text parts identified as NEs are then replaced by NE labels, resulting in anonymised text and (if needed) a mapping table. The sentences may be manually reviewed. Finally, the sentences can be recombined into paragraphs and the layout added to them, so the resulting document can be archived or shared with other organisations.

Another basic scenario consists of a workflow for processing documents with high sensitive data; this workflow proceeds in a similar fashion as the above one, but involves replacement of (some) NEs by other NEs in order to reduce the risk of deidentification.

A third basic scenario consists of anonymising translation memories, by checking source and target sentences separately or in combination, using some type of replacement strategy (e.g. NE labels).

One more advanced scenario consists of training an application from anonymised monolingual or bilingual data, such as an MT system or a summariser. This may

[14] Nonetheless, the substitution of an NE by another one may also lead to a result that reveals what has been replaced. For instance, in morphologically complex languages, the inflection of the replacing NE may not fit the rest of the sentence.

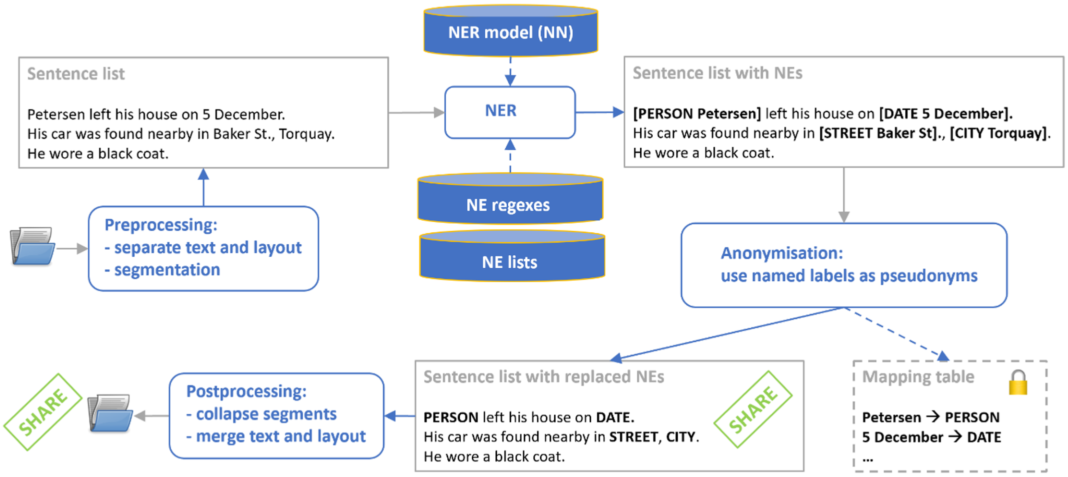

NER model (NN)
Sentence list
Petersen left his house on 5 December.
His car was found nearby in Baker St., Torquay.
He wore a black coat.
NER
Sentence list with NEs
[PERSON Petersen] left his house on [DATE 5 December].
His car was found nearby in [STREET Baker St]., [CITY Torquay].
He wore a black coat.
NE regexes
NE lists
Preprocessing:
- separate text and layout
- segmentation
Anonymisation:
use named labels as pseudonyms
SHARE
Postprocessing:
- collapse segments
- merge text and layout
Sentence list with replaced NEs
PERSON left his house on DATE.
His car was found nearby in STREET, CITY.
He wore a black coat.
SHARE
Mapping table
Petersen → PERSON
5 December → DATE
...

Figure 1: Scenario for workflow using NE labels

require customised ways of annotating training data and input (e.g. the inclusion of other annotations than just the NE labels, such as linguistic codes) and, possibly, postprocessing the system output (deanonymisation).

Another advanced scenario consists of retraining an NER model. While the above scenarios make use of a fixed model, the results of which can be complemented with user-specified lists and patterns, they could be extended with components for retraining the NER model based on the manual correction of the NER results (i.e. finetuning the model). In case anonymisation should take place on highly domain specific text (e.g. in the medical domain), it may even be necessary to make use of a non-generic NER model with an appropriate set of NE labels, or even to train such a model from scratch based on sentences manually annotated with labels.

3.5. Proof-of-concept software

In order to demonstrate the basic scenarios described in the previous section, we set up proof-of-concept software in the form of docker images and a user interface (UI) addressing those. The docker images allow for creating lightweight, standalone virtual machines that include everything needed to run an application, thus facilitating setup of the proof-of-concept software for potential users and allowing them to run it in their own environment, thus providing full control. The docker images integrate the three publicly available tools described in Section 3.2. They also add new functionality to them: support for document formats like docx, for formats storing bilingual data, for applying user-specified NE lists and patterns, for viewing the mapping table, and for obtaining a detailed logging of the commands run in the background. The dockers and the UI will be provided on a GitHub repository.

Figure 2 shows the components of the UI. It allows for selecting a tool to try out: the TM-Anonymizer (in monolingual or bilingual mode), the COMPRISE Text Transformer (monolingual), or the Biroamer (bilingual). In case of the monolingual mode of the TM-Anonymizer, the user can select a language, a docx file and, optionally, a file containing NEs and/or patterns. The resulting anonymised text contains NE labels. In case of the bilingual mode, the user specifies the source and target language and a TMX file.[15] The COMPRISE Text Transformer tab in the UI also takes a docx file as input, and allows for specifying the type of replacement: replacement with label *X* or replacement with another NE.[16] Finally, the Biroamer tab allows for anonymising a TMX file by replacing NEs with their labels, and (optionally) shuffling translation units and omitting a fraction of them. In sum, the UI allows a potential user to find out the opportunities and challenges of anonymisation tools in their environment.

4. Multilingual Fake News Processing

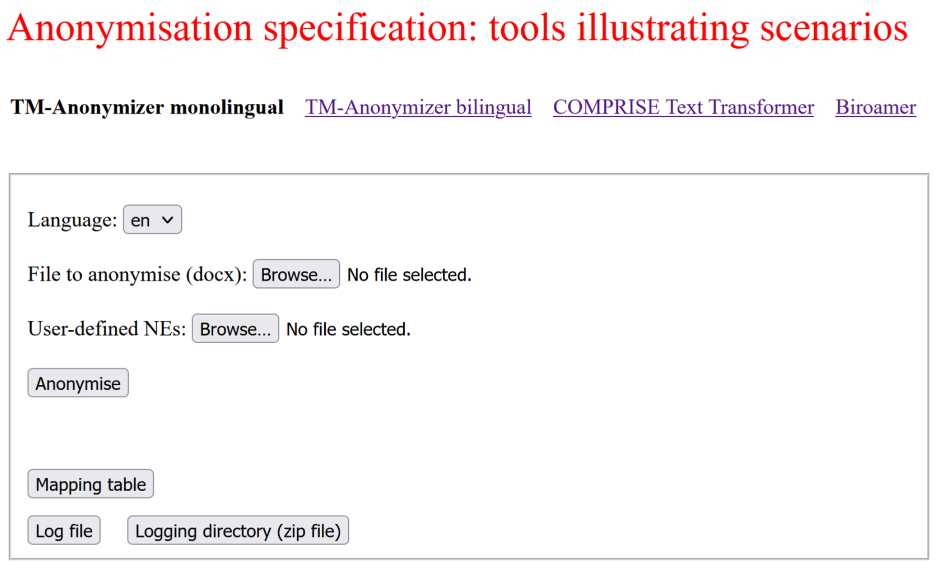
In this section, we motivate ELRC's specification for multilingual fake news processing. The specification being work in progress, we provide a general description of the relevant tools and techniques, the consultation round, and the experiments that we are performing in parallel with this round and will lead to proof-of-concept software in the course of 2022.

4.1. Motivation

In an increasingly digital world, disinformation (fake news) on various topics is spreading quickly, at a much faster rate than accurate information. Disinformation consists of false information with the intent to manipu-

[15] Translation Memory eXchange, an XML specification for the exchange of translation memories.

[16] In the second case, the user may choose to replace multiword NEs by other multiword NEs or to replace each word in the NE separately.



Anonymisation specification: tools illustrating scenarios

TM-Anonymizer monolingual TM-Anonymizer bilingual COMPRISE Text Transformer Biroamer

Language: en

File to anonymise (docx): No file selected.

User-defined NEs: No file selected.

Figure 2: User interface of proof-of-concept anonymisation software

late, spread and cause harm. Social media are the primary platform for this spread.

Disinformation has known a huge boost during the US presidential election of 2016 (Allcot and Gentzkow, 2017). This resulted in fake news acquiring global prominence, which fueled interest in this topic within the academic community. A further boost has been given as a result of the COVID-19 pandemic (Alam et al., 2021). The World Economic Forum ranks the spread of fake news as among the world's top global risks (World Economic Forum, 2018). It leads to economic, social and political damage and can manipulate public opinions and fuel interpersonal conflicts between individuals or groups of individuals.

Because of the quick spreading rate of disinformation, the availability of tools for automatically detecting it and making readers aware of the potential harm of this content is gaining urgency. Moreover, as disinformation is a global phenomenon, it is important to explore to what extent multilingual coverage has been or may be reached in the area of fake news processing. This multilingual aspect is explored as part of the ELRC action.

4.2. Tools and techniques

Fake news detection consists of classifying an article as true (authentic) news or fake news, or using some other distinction, such as positive/negative. It should be noted in this respect that the fake news detection task should not be confused with fact-checking, a closely related task in which an expert or journalist is asked to judge whether a specific claim in a news article or social media post is correct based on evidence. Specific methodologies and model architectures exist for this more challenging task; see for example the FEVER (Fact Extraction and VERification) shared task (Thorne et al., 2018). Nonetheless, models for fake news detection may also be helpful in order to speed up the process of fact-checking by flagging articles potentially containing false claims. Therefore, fake news detection should be considered as a mechanism for determining whether an article is likely to have the intent to deceive, e.g. by using language that aims at triggering an emotional response from the reader. In this respect, a classifier can be considered as a means to support human fact-checkers, who perform an in-depth verification of the items classified as fake news.

Several architectures for fake news detection were proposed in recent years, following the advances made in the more general field of text classification. For instance, a model may be trained on linguistic features extracted from news articles (Patwa et al., 2021; Pérez-Rosas et al., 2018; Ahmed et al., 2017), or fake news classification may be based on pretrained word vectors and Convolutional Neural Networks (Zhou et al., 2020). Similarly as for many text classification tasks, the use of BERT (Bidirectional Encoder Representations from Transformers) language models (Devlin et al., 2019) has shown to be very effective for detecting fake news; see for instance Kaliyar et al. (2021).

The majority of machine learning models used for fake news classification are language-agnostic, i.e. the performance of such models only depends on the training data, whichever language they are written in. As mentioned above, a pre-trained BERT model, possibly multilingual, can be used as a basis for fine-tuning towards the classification task. Publicly available datasets have a limited scope in terms of languages (mostly English), despite the global character of the disinformation phenomenon. This shows the need to spend effort to support more languages.

Besides the features inherent in an article's text, the

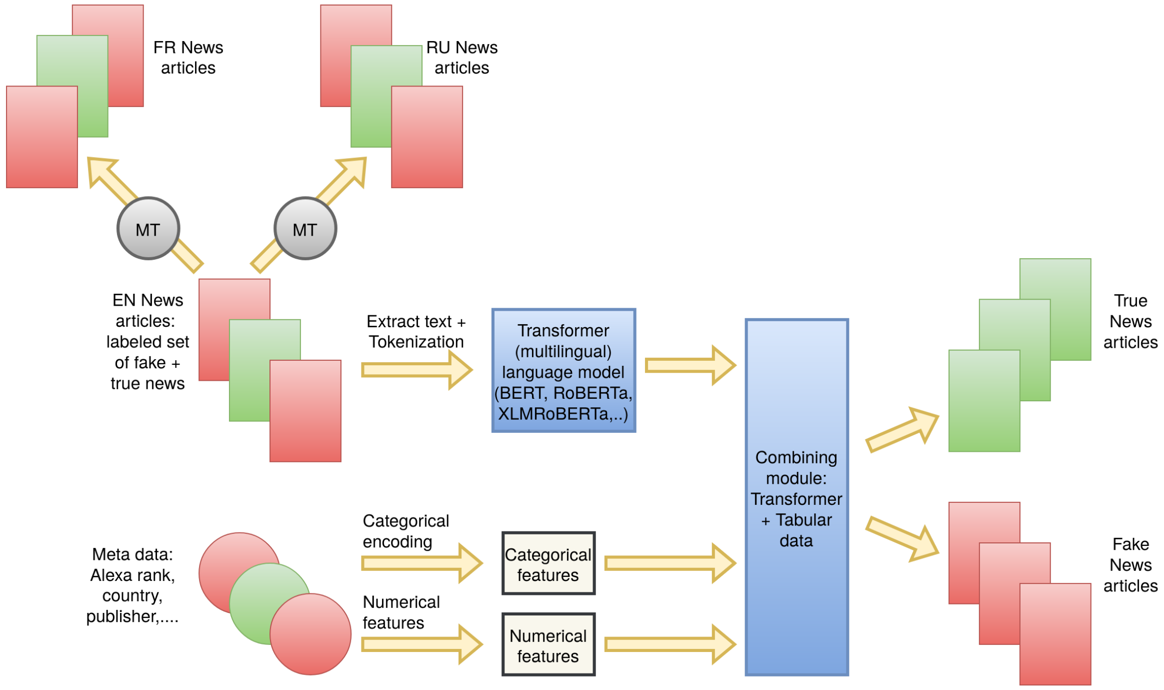

FR News articles
RU News articles
MT
MT
EN News articles: labeled set of fake + true news
Extract text + Tokenization
Transformer (multilingual) language model (BERT, RoBERTa, XLMRoBERTa,..)
Combining module: Transformer + Tabular data
True News articles
Fake News articles
Meta data: Alexa rank, country, publisher,....
Categorical encoding
Categorical features
Numerical features
Numerical features

Figure 3: Supervised approach to fake news detection

source of an article can also be used as a feature during classification, some sources being ill-reputed. Another feature that can be used is the cross-lingual presence of the information conveyed in an article: using an MT system, the information can be compared to that conveyed in articles written in other languages[17] (Dementieva and Panchenko, 2021). The more cross-lingual evidence is found for an article, the more likely the information in the article constitutes true news; the rationale is that fake news receives less response across the global media than genuine news.

4.3. Consultation round

In the consultation round, stakeholders with various profiles are being consulted. A major stakeholder is the European Digital Media Observatory (EDMO),[18] a partnership between various organisations that provides a collaboration platform bringing together fact-checkers, media literacy experts, journalists, academic researchers, media organisations, etc. EDMO cooperates with eight national hubs. Other meetings in the consultation round involved stakeholders from academia and industry working on automated approaches to fake news detection, for instance from Skolkovo Institute of Science and Technology (Moscow). In addition, fact-checking organisations are being approached, in order to shed light on the procedures they apply for manual fact-checking and the tools they make use of.

4.4. Experiments

A major stumbling block when building and testing fake news detection systems is the lack of training data. Publicly available datasets with articles labeled as fake or true news are not only restricted in number of languages, as mentioned earlier, but also in terms of domain and time coverage. As topics dealt with in fake news quickly evolve, existing training sets may be of little help to classify more recent articles. However, collecting relevant true news articles can be relatively easy, for example by making the assumption that large news agencies and renowned newspapers publish articles that can generally be considered authentic. Therefore, we also propose a (multilingual) methodology for detection of fake news that uses models only trained on news articles labeled as true. We will refer to this methodology as unsupervised fake news detection.

Concerning supervised classification, we use an approach in which features inherent in the text are combined with other features. Our classifier takes as input the text itself as well as categorical features (e.g. “publisher”) and numerical features (e.g. Alexa rank, which is a measure of website popularity), and predicts whether the data point (article) is true or fake news. Due to the lack of large multilingual datasets for fake news processing, we propose the use of MT for obtaining a multilingual model, i.e. an English (EN) (sub)set of labeled news articles is translated to a language of choice, e.g. French (FR) or Russian (RU), after which the classification model is trained on the concatenation of the EN, FR and RU training sets. This is illustrated in Figure 3.

[17] By comparing sentence embeddings through cosine similarity.

[18] https://edmo.eu

As for unsupervised classification, we have set up a novel approach which aims at training a model merely on articles known to constitute true news, in order to tackle data sparsity. This approach makes use of a technique called one-class classification, or anomaly detection, which is already being used in the field of computer vision and natural language processing (Schlegl et al., 2017; Ruff et al., 2019). We applied an existing algorithm called Context Vector Data Description (CVDD) (Ruff et al., 2019) in a multilingual way, by aligning vector representations. By training a One-Class Support Vector Machines (OC-SVM) model (Scholkopf et al., 2011) on the obtained anomaly score and a set of features extracted from news articles related to punctuation, readability and syntax, we build an anomaly detection model that is merely trained on true news articles. At inference time, articles which are anomalies according to the model are considered fake news.

When evaluating our proposed architecture for the supervised (multilingual) models, making use of the CEF eTranslation service[19] for MT, we observed that good classification results can be achieved when training and evaluating on EN, FR and RU data. Moreover, in line with Casula and Tonelli (2020), we observed that our models achieve similarly good results in a zero-shot setting (i.e. training on language A and evaluating on language B). For the unsupervised models, our current observations indicate that, despite the training process only making use of true news articles, promising performance can be achieved. Interestingly, we also observe zero-shot potential for these type of models, especially for closely related languages.

A limitation of the proposed approach for both the supervised an unsupervised model architectures is that, due to the lack of large multilingual datasets for fake news detection, we use MT to obtain multilingual datasets. Although this is a common approach for obtaining cross-lingual models, text that is machine-translated from language A to language B will typically be closer to language A than text originally written in language B. Therefore, the cross-lingual aspect of the classification task may be more complex for an in vivo scenario.

5. Conclusions

In this paper, we described the LT assessments performed as part of the ELRC action of the EC, which aims towards minimising language barriers in the EU. The assessments consist of testing various tools and techniques, documenting them in a hands-on way, performing experiments with them, and setting up proof-of-concept environments that demonstrate their potential and their challenges to EC staff and EU Member State representatives, thus facilitating their uptake by public sector users. We zoomed in on the two most extensive assessments (LT specifications), involving a consultation round with various types of stakeholders.

In the Automated Anonymisation specification, which is in its final steps, we investigated tools and techniques for deidentifying monolingual or bilingual text. They replace NEs and specific patterns with NE labels or with other words of a similar type, thus supporting the effort to make text GDPR compliant for organisations that want to store text containing personal data and to share this text with other organisations. The consultation round has shown that the sensitivity of the information has an influence on the choice of replacement strategy; the use of similar words instead of NE labels is better suited for hampering malicious attempts at reidentification, but also more challenging from a linguistic point of view. The consultation round has also shown that a lot of exploration is still needed for constructing MT systems able to translate anonymised text, and that the user of an anonymisation tool needs to have sufficient control, i.e. the possibility to create custom NE lists and patterns and, ideally, to run the tool in-house.

In the Multilingual Fake News Processing specification, which is ongoing work, we investigated tools and techniques for detecting articles that spread false information in order to deceive readers and, as such, provide damage on political or other levels. Despite the global character of disinformation, the publicly available datasets required for training deep learning models are limited in terms of languages. Therefore, the specification zooms in on ways to increase multilingual support. We are experimenting with supervised classification, by making use of not only text-inherent but also categorical and numerical features, such as the Alexa rank. We are also experimenting with a novel approach for unsupervised classification: we apply the technique of anomaly detection and train a model that also uses various types of features but only makes use of articles known to constitute true news. This strategy aims at reducing the impact of data sparsity, on the level of language as well as topics. When applying the model to an unseen article, it is considered fake news if it is an anomaly according to the model.

6. Acknowledgements

The work described in this paper is funded by the SMART 2019/1083 service contract (running from 2020 to 2022), one of the service contracts through which the ELRC action is carried out. The contracting authority is the EC's DG CNECT, Unit Multilingualism (G3).

7. Bibliographical References

Abadeer, M. (2020). Assessment of DistilBERT performance on named entity recognition task for the detection of protected health information and medical concepts. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 158–

[19] https://webgate.ec.europa.eu/etranslation

167, Online, November. Association for Computational Linguistics.

Adelani, D. I., Davody, A., Kleinbauer, T., and Klakow, D. (2020). Privacy guarantees for de-identifying text transformations. In Helen Meng, et al., editors, *Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020*, pages 4666–4670. ISCA.

Ahmed, H., Traore, I., and Saad, S. (2017). Detection of online fake news using N-Gram analysis and machine learning techniques. In *International Conference on Intelligent, Secure and Dependable Systems in Distributed and Cloud Environments*, pages 127–138.

Alam, F., Dalvi, F., Shaar, S., Durrani, N., Mubarak, H., Nikolov, A., Da San Martino, G., Abdelali, A., Sajjad, H., Darwish, K., and Nakov, P. (2021). Fighting the COVID-19 infodemic in social media: a holistic perspective and a call to arms. In *Proceedings of the Fiftheenth International AAAI Conference on Web and Social Media*, pages 913–922.

Allcot, H. and Gentzkow, M. (2017). Social media and fake news in the 2016 election. In *Journal of economic perspectives*, pages 211–236.

Bañón, M., Chen, P., Haddow, B., Heafield, K., Hoang, H., Esplà-Gomis, M., Forcada, M. L., Kamran, A., Kirefu, F., Koehn, P., Ortiz Rojas, S., Pla Sempere, L., Ramírez-Sánchez, G., Sarrías, E., Strelec, M., Thompson, B., Waites, W., Wiggins, D., and Zaragoza, J. (2020). ParaCrawl: Web-scale acquisition of parallel corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4555–4567, Online, July. Association for Computational Linguistics.

Casula, C. and Tonelli, S. (2020). Hate speech detection with machine-translated data: the role of annotation scheme, class imbalance and undersampling. In *Proceedings of the Seventh Italian Conference on Computational Linguistics*.

Dementieva, D. and Panchenko, A. (2021). Cross-lingual evidence improves monolingual fake news detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 310–320, Online, August. Association for Computational Linguistics.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Dwork, C. (2006). Differential privacy. In Michele Bugliesi, et al., editors, *Automata, Languages and Programming, 33rd International Colloquium, ICALP 2006, Venice, Italy, July 10-14, 2006, Proceedings, Part II*, volume 4052 of *Lecture Notes in Computer Science*, pages 1–12. Springer.

European Union Agency for Cybersecurity. (2019). Pseudonymisation techniques and best practices. Technical report, November 2019.

Kaliyar, R. K., Goswami, A., and Narang, P. (2021). FakeBERT: Fake news detection in social media with a BERT-based deep learning approach. In *Multimedia Tools and Applications*, pages 80:11765–11788.

Kamran, A., Dzeguze, D., van der Meer, J., Panic, M., Cattelan, A., Patrioli, D., Bentivogli, L., and Turchi, M. (2020). CEF Data Marketplace: Powering a long-term supply of language data. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 459–460, Lisboa, Portugal, November. European Association for Machine Translation.

Lösch, A., Mapelli, V., Choukri, K., Giagkou, M., Piperidis, S., Prokopidis, P., Papavasiliou, V., Deligiannis, M., Berzins, A., Vasiljevs, A., Schnur, E., Declerck, T., and van Genabith, J. (2021). Collection and curation of language data within the European Language Resource Coordination (ELRC). In *Proceedings of the QURATOR 2021 Conference on Digital Curation Technologies*.

Patwa, P., Sharma, S., Pykl, S., Guptha, V., Kumari, G., Akhtar, M. S., Ekbal, A., Das, A., and Chakraborty, T. (2021). Fighting and infodemic: COVID-19 fake news dataset. In *International Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situation*, pages 21–29.

Pérez-Rosas, V., Kleinberg, B., Lefevre, A., and Mihalcea, R. (2018). Automatic detection of fake news. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3391–3401, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

Ruff, L., Zemlyanskiy, Y., Vandermeulen, R., Schnake, T., and Kloft, M. (2019). Self-attentive, multi-context one-class classification for unsupervised anomaly detection on text. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4061–4071, Florence, Italy, July. Association for Computational Linguistics.

Schlegl, T., Seeböck, P., Waldstein, S. M., Schmidt-Erfurth, U., and Langs, G. (2017). Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In *International conference on information processing in medical imaging*, pages 146–157. Springer.

Scholkopf, B., Platt, J. C., Shawe-Taylor, J., Smola, A. J., and Williamson, R. C. (2011). Estimating the support of a high-dimensional distribution. In *Neu-*

ral Computation, pages 13(7):1443–1471.

Smal, L., Lösch, A., van Genabith, J., Giagkou, M., Declerck, T., and Busemann, S. (2020). Language data sharing in European public services – overcoming obstacles and creating sustainable data sharing infrastructures. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3443–3448, Marseille, France, May. European Language Resources Association.

Stanislawek, T., Wróblewska, A., Wójcicka, A., Ziembicki, D., and Biecek, P. (2019). Named entity recognition - is there a glass ceiling? In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 624–633, Hong Kong, China, November. Association for Computational Linguistics.

Sweeney, L. (2000). Simple demographics often identify people uniquely. *Health (San Francisco)*, 671:1–34.

Thorne, J., Vlachos, A., Christodoulopoulos, C., and Mittal, A. (2018). FEVER: a large scale dataset for fact extraction and verification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 809–819.

World Economic Forum. (2018). The global risks report 2018. Insight Report. 13th edition.

Zhou, X., Mulay, A., Ferrara, E., and Zafarani, R. (2020). Recovery: A multimodal repository for COVID-19 news credibility research. In *The 29th ACM International Conference on Information and Knowledge Management*, pages 3205–3212.