ER documents categorization and explanation

Enrico Mensa
Computer Science Department, University of Turin
Davide Colla
Computer Science Department, University of Turin
Marco Dalmasso
Servizio sovrazonale di Epidemiologia, ASL TO3 Regione Piemonte
Marco Giustini
Reparto di Epidemiologia Ambientale e sociale, Dipartimento Ambiente e Salute (DAMSA), Istituto Superiore di Sanità
Carlo Mamo
Servizio sovrazonale di Epidemiologia, ASL TO3 Regione Piemonte
Alessio Pitidis
Reparto Epidemiologia ambientale e sociale Dipartimento Ambiente e Salute (DAMSA)
DANIELE PAOLO RADICIONI (daniele.radicioni@unito.it)
University of Turin https://orcid.org/0000-0003-0443-7720

Research article

Keywords: XAI, explanation, text categorization, categorization explanation, word embeddings, semantic frames, slot filling, event extraction, violent event tracking

Posted Date: January 16th, 2020

DOI: https://doi.org/10.21203/rs.2.21065/v1

Version of Record: A version of this preprint was published on October 15th, 2020. See the published version at https://doi.org/10.1186/s12911-020-01237-4.

RESEARCH

ER documents categorization and explanation

Enrico Mensa[1†], Davide Colla[1†], Marco Dalmasso[2], Marco Giustini[3], Carlo Mamo[2], Alessio Pitidis[3] and Daniele P Radicioni[1*†]

Abstract

Background: Emergency room reports are a specific kind of text, posing specific challenges to natural language processing techniques. In this setting, violence episodes on women, elderly and children are often under-reported. Categorizing textual descriptions as containing violence-related injuries vs. non-violence-related injuries, is thus a relevant task, to the ends of devising alerting mechanisms to track violence episodes.

Methods: We present a system to detect episodes of violence from the textual descriptions contained in emergency room reports. It employs a deep neural network for categorizing textual ER reports data. Additionally, the system complements such output by making explicit which elements corroborate the interpretation of the record as reporting about violence-related injuries. To these ends we designed a novel hybrid technique for filling semantic frames that employs distributed representations of the terms herein, along with syntactic and semantic information.

Results: We tested our system on a set of real data of emergency room reports, coming from an Italian branch of the EU-Injury Database (EU-IDB) project, annotated by hospital staff. Our experimentation shows that the system produces accurate categorization (of violent vs. non violent records), paired with interesting results on the explanation of such output. At times, it also allowed unveiling annotation errors committed by hospital staff.

Conclusions: In the last few years deep architectures and word embeddings have been compared to a tsunami hitting AI and the area concerned with natural language processing. Only at a later time we have been realizing that the stunning output of deep networks needed to be explained: our proposal, combining distributed and symbolic (frame-like) representations are a possible answer to this pressing request for interpretability. Although the present application is focused on the medical domain, the proposed methodology is general and, in principle, it can be extended to further application areas and categorization tasks.

Keywords: XAI; explanation; text categorization; categorization explanation; word embeddings; semantic frames; slot filling; event extraction; violent event tracking

Introduction

Explanation is acknowledged to be an epistemologically relevant process [1] and a precious feature to build robust and informative systems. It is a matter of fact that artificial explanation has a long tradition in the AI field, and some areas such as case-based reasoning seem to be intrinsically connected to explanatory needs [2, 3]. In machine learning, decision trees [4] and sparse linear models [5] are popular examples of techniques that produce interpretable models. Many sorts of explanation can be drawn, responding to diverse needs underlying the general aim at providing more transparency to algorithms and systems. For example, the role of explanation in AI systems and its relevance w.r.t. systems accountability is debated in the EU General Data Protection Regulation [6, 7]. On a different side, the tight relation between automatic explanation and trust has been individuated in many contexts as a central issue (think, e.g., to the role of explanation in the field of information security), in its interplay with ethical and sociological issues [8]. Besides, together with the impetuous surge of work on explainable AI, some attempts have been carried out at investigating what constitutes a good explanation, and how the contributions from different disciplines such as psychology and cognitive science can enrich the quality of the explanations being provided by systems [9].

*Correspondence: daniele.radicioni@unito.it
[1]Department of Computer Science, University of Turin, Corso Svizzera 185, 10149 Turin, ITALY
Full list of author information is available at the end of the article
†Equal contributor

Areas where intelligent systems and agents are currently deployed are as different as personal assistants, logistics, scientific research, law and health care. While in some cases (such as, e.g., some kinds of personal assistants) users are not interested in explanations, for sensitive tasks involving "critical infrastructures and affecting human well-being or health, it is crucial to limit the possibility of improper, non-robust, and unsafe decisions and actions" [10]. One chief motivation for building explainable AI systems is thus the need to check systems behavior, to ensure that systems perform as expected. This need has become particularly relevant in the last few years, that have witnessed the spread of deep learning based neural networks, in that these are featured by strong predictive power, at the expense of the interpretability of their output [11, 12]. In this work we investigate one such critical application domain: the categorization of electronic medical records (EMR) data, where an Information Extraction approach has been devised to complement the output of the deep neural network performing the categorization step.

In particular, our system is aimed at categorizing textual descriptions from Emergency Room Reports (ERRs) as containing violence-related injuries vs. non-violence-related injuries, to the ends of devising alerting mechanisms to track violence episodes. The early detection of violence in general, and specifically against women, elderly and childhood is a serious concern for our societies. However, interestingly enough, such phenomena are to date underestimated and not even fully recorded in statistics. Let us consider, for example, that violence against women is seldom reported from its inception due to many reasons, such as the fact that this sort of violence is performed by family members or acquaintances [13, 14]. Likewise, and due to similar reasons, according to the recommendations by Centers for Disease Control and Prevention (CDC), violence on children is largely acknowledged to be under-reported [15]. Additionally, hospital staff may have practical difficulties in properly annotating violence episodes (e.g., complex user interfaces, or lack of time to fully describe the medical history of patients), so that violence and its effects are to date not fully grasped. This determines the necessity to devise automatic systems to automatically detect violence in electronic medical records (EMR) data, so to allow timely intervention and design policies to contrast the phenomenon of violence. From a technical viewpoint, a desideratum would be that of building a classifier to individuate EMRs containing descriptions of violent events in the medical history along with its effects in the physical examination. In order to generate an explanation of the obtained categorization, one would also be able to make explicit the more relevant elements associated to events: by whom the violence was exercised, in what ways, what trauma was produced on the victim, which are the involved body parts, when and where the event has occurred.

This is the focus of the work: we face the problem of extracting meaningful pieces of information to the ends of justifying the categorization performed by the system. We present the VIDES system, so dubbed after 'VIOLENCE DETECTION SYSTEM': the designed approach provides violence events with a formal characterization in terms of semantic frames [16]. Additionally, the control strategy devised to fill the frame slots employs a hybrid strategy exploiting distributed word representations together with morphological (on part-of-speech tags) and semantic (on super-senses) information.

Related Work

The closely related task of frame identification has been addressed by [17]: in this work distributed representations of predicates and their syntactic context were exploited, paired with a general purpose set of word embeddings. Our work differs from the mentioned approach, in that we do not make use of syntactic information (since our input is very noisy, which would completely undermine parsing accuracy and reliability). Additionally, we retrain our embeddings on a set of EMR data, to acquire specific descriptors (we are concerned with a very specific application domain, that of first aid medical records) for the Italian language; and finally we are concerned with a more restricted task, that is extracting the fillers for the slots from a single frame, the violence frame.

As regards as acquiring distributed representations to describe verb dependents and semantic frames, word embeddings have been employed also to investigate cross-language misalignment, such as related to polysemy, syntactic valency (i.e., the number of dependent arguments of verbs), and lexicalization [18]. In particular, the authors of the cited work build different embeddings for a given frame, one for each language of interest. Since such embeddings lie in the same semantic space, this approach is used to automatically measure the cross-lingual similarity of language-specific frames to the ends of investigating the possibility of frame transfers across languages.

Word embeddings have been used also to perform semantic role labeling (SRL); this task is to discover the relations between between predicate and its arguments, so it basically amounts to discovering "who" did "what" to "whom", "when", "where", and "how". This line of research was started in [19], where the distributions over verb-object co-occurrence clusters

were used to improve coverage in argument classification. The work by [20] proposes a distributional approach for acquiring a semi-supervised model of argument classification preferences, that is used to reduce the complexity of the employed grammatical features in combination with a distributional representation of lexical features. Additionally, in [21] a selectional preference model has been proposed providing a single additional feature to classify potential arguments based on distributional similarity. The neural network architecture described in [22] relies on the intuition of embedding dependency structures, and jointly learns embeddings for dependency paths and feature combination. The work by [23] proposes to tackle the SRL task by assigning semantic roles through a feedforward network that uses a convolution function over windows of words; interestingly enough, this system does not make use of syntactic information.

With respect to this line of research using word embeddings to perform the SRL task, we face a slightly different problem. First, we are not really concerned with SRL: we are interested in a variant of SRL, where we need to extract salient information (to generate an explanation) associated to a single semantic frame (describing violent events). Additionally, different from the surveyed approaches, our input texts are very challenging and cannot undergo a standard parsing process, as almost any sentence contains typos, acronyms, domain-specific (at times, hospital-specific) abbreviations, and clauses well-formedness is mostly violated. Such features prevented from designing a suitable sequence of preprocessing steps, and our system deals with all mentioned phenomena without performing rewriting of the input text. This implies that our system substantially differs from those concerned with the SRL task. In fact, most SRL modules perform two main steps, argument identification and argument classification, with the former basically grounded on syntactic parsing, and the latter requiring additional semantic knowledge to solve the task. Instead, our approach puts together word embeddings, supersense tags, and simple part-of-speech (PoS) filtering techniques to the ends of collecting enough information to explain why an Emergency Room Report describes a violence event.

The System

The developed system relies on two main modules. The first module performs the classification of the medical records in order to assign a label, determining whether the record exhibits traits of violence or not. The second module, on the other hand, is aimed at extracting salient information from the violent record by adopting a hybrid approach that exploits distributional, semantic and syntactic information.

The Neural Model for the Categorization Step

As regards as the categorization of the medical records, a neural model has been devised to discriminate among violent and non-violent entries. Input to the model is the text contained in the ER record; such text is first tokenized and mapped onto a numerical vectorial representation. The mapping from terms to vectors $\langle term, numerical\ id \rangle$ was acquired from the considered dataset. More specifically, the weight matrix has been initialized with 300-d FastText word vectors trained on the whole dataset by adopting the Skip-Gram architecture [24]. The input layer is then connected to a single dimension convolutional layer, which is composed by 64 filters; the kernel consists of 5 units and adopts the Rectified Linear Unit (ReLU) activation function. A dropout rate of 20% was set between the input layer and the convolutional layer in order to reduce the overfitting of the model. A max pooling layer with a window size set to 4 units was adopted to reduce the dimension of the input, and is followed by an LSTM layer built with 100 memory units. Finally, a fully connected layer —adopting the sigmoid activation function— is used to predict the class of the the medical record: V for violent episodes, and NV for non-violent episodes. In this setting the role of the convolutional layer is twofold: *(i)* to learn abstract features coming from medical reports; and *(ii)* to reduce the training time. The whole architecture is illustrated in Figure 2. The training phase employs Adam stochastic optimization [25] and binary cross entropy loss function.

Building Explanations by Frame Elements Embedding

The second module is fed with the entries that were classified as violent by the network, and is intended to extract information relevant to describe a violence episode: this amounts to filling the slots (that can be though of as object fields) of a violence frame. The violence frame contains the most salient information ordinarily associated to violence events, and it is thus defined as follows.

- AGENT: The agent performing the violence. Examples phrases may be 'known person', 'husband', 'wife', etc.;
- MODE-INSTRUMENT: The mode or the instrument adopted while performing the violence. Examples of this field are 'punch', 'aggression', 'knife', etc.;
- TIME: Temporal information regarding when the violence occurred. Examples are 'evening', 'night', 'today', etc.;
- LOCATION: The physical place in which the violence took place. Examples are 'home', 'workplace', 'bus station', etc.;
- BODY-PART: Body part harmed by the violent act. Examples are 'arm', 'head', etc.;

- LESION-TYPE: Type of injury produced by the violent act. Examples are 'fracture', 'contusion', 'trauma', etc..

All of the mentioned fields may have zero or multiple fillers, depending on the content of the considered entry. Also, attached to each field f we have two lists: PoS_f and SST_f, indicating the part-of-speech (PoS) tags and SuperSense tags (SST) that a filler for f can assume. The supersense tagging consists of annotating text with the tagset defined by the 41 WordNet [26] super-sense classes for nouns and verbs. Such top elements define broad semantic categories, such as `SST.NOUN_PERSON` or `SST.NOUN_LOCATION`, that may be relevant to fill our frame slots or to rule out some elements. Since the tagset is directly related to WordNet synsets, this information can be intended as a partial word sense disambiguation [27].

Such information is subsequently used to match the semantic need of the frame with the morphological and semantic information available in the actual input text. For instance, the AGENT field can only be filled by a `SST.NOUN_PERSON` and `POS.NOUN`.

As mentioned, input texts from the ER records are quite noisy, to such an extent that only few records can be found where all words are in a standard dictionary, thereby determining the need for a preprocessing step in which the text is cleaned and normalized. In order to perform the extraction we take into consideration the sentences attached to the entry and remove all punctuation. We then identify locutions (which we call *extended tokens*) whose elements are common multiword expressions found in the dataset that we would like to process as single tokens (e.g., *known_person*). Finally, the sentences are tokenized.

We then proceed to the construction of a candidate set of fillers for each field: given a field f, we initialize its set of candidates C to all terms. Then, we prune from C all terms whose PoS(s) or SST(s) are not compatible with the needs of the semantic field f. More precisely, for each term $t \in C$ we retrieve its PoS and its most frequent SSTs. Namely, PoS tagging is computed through the Tint parser, which is a porting for the Italian language [28] of the Neural Network Dependency Parser [29]. Supersense tags are computed by accessing WordNet and retrieving the most frequent sense, among all senses possibly underlying a given input term. Although this may seem a too crude simplification, the most frequent sense is experimentally acknowledged as a competitive baseline [27], and used as a core feature in more sophisticated SST systems [30], that ensures limited computation time and effort. Given the rather narrow semantic domain for the present application, we opted for this simple heuristics. The term t is retained only if its PoS is included in PoS_f and at least one of its SSTs is included in SST_f. Extended tokens bypass this process, and they are all included as candidates by default.

Once we have filtered C so that it contains only terms that are allowed as fillers for f, we rank them by leveraging the FastText embeddings that were acquired by the first module. Namely, for each field we build a synthetic vector by averaging the most frequent terms that could act as filler for the given field. The similarity between this vector and all candidates in C is then computed, so that the candidates can be ranked.

The last phase of the algorithm consists in building the final answer provided by the system. Here, we apply two strategies: *(1)* all the candidates that have a similarity lower than a certain threshold are discarded; and *(2)* if a term is a candidate for more than one field, it is assigned to the field in which it appears with maximum similarity.

Figure 1 provides an example that has been translated from Italian into English to illustrate the whole process.

Evaluation

Dataset and Procedure

The data used in the experimentation are real-world emergency room reports (ERRs) collected in Italian Hospitals, and then made available by the Italian National Institute of Health in the frame of the SINIACA project [31]. The SINIACA project (so dubbed after 'Sistema Informativo Nazionale sugli Incidenti in Ambiente di Civile Abitazione', National Information System on Accidents in Civil Housing Environment) is the Italian branch of the European Injury Database (EU-IDB, `https://ec.europa.eu/health/indicators_data/idb_en`), an EU-wide surveillance system concerned with accidents, collecting data from hospital emergency department patients according to EU recommendation. SINIACA is a data collection on home injuries, based on a sample of hospital emergency departments, in implementation of the recommendation of the Council of the European Union no. C 164/2007/01 on injury prevention and safety promotion.

For our experimentation we have used 153,823 records from the SINIACA-IDB, which were originally annotated by hospital staff as containing injuries descending from either violent (V in the following) or non violent acts (NV in the following). The dataset is very unbalanced, as it contains 5,168 records that were tagged as violent, while the remaining 148,655 (96.64%) were labeled as not caused by violent acts. The dataset has been randomly split into 2 equal parts: the former one was used for training and parameters

tuning (80:20 the ratio between training and validation set, respectively); the rest was used as our test set. The partitioning was managed by preserving the distribution of V/NV items: namely, we maintained the same ratio between violent and non-violent entries as occurring in the considered data, where the 3.36% of the entire dataset belongs to the violent class.

As regards as the two modules of our system, we have then recorded the classification accuracy obtained by the classifier implemented through the neural network. As regards as the evaluation of the explanation generated, we annotated 200 randomly sampled records among those returned as violent (V) at categorization time. Each such record was associated to a frame, whose fields were filled with the information available in the text document. Provided that each frame contains 6 fields, overall 1200 slots were annotated: in 729 cases a filler was annotated from the accompanying text, whilst in 471 cases no value could be set. More specifically, the available information associated widely varied across the slots, as follows: AGENT was filled in the 60% of cases; MODE-INSTRUMENT was filled in the 97% of cases; TIME was filled in the 23.5% of cases; LOCATION was filled in the 8% of cases; BODY-PART was filled in the 89.5% of cases; LESION-TYPE was filled in the 86.5% of cases. Since multiple annotations were allowed (according to the information available in the input text), we recorded overall 5.53 fillers annotated for each record (e.g., the lesion type can be both 'trauma' and 'wound'; the involved body part can be 'shoulder', 'leg' and 'arm'): more specifically AGENT was filled on average with 0.66 elements; MODE-INSTRUMENT was filled on average with 1.44 elements; TIME was filled on average with 0.28 elements; LOCATION was filled on average with 0.09 elements; BODY-PART was filled on average with 1.77 elements; LESION-TYPE was filled on average with 1.29 elements.

Such annotated data was set as our ground truth annotation, against which the frame computed by the explanation module was compared.

Results

Categorization Results

The categorization is aimed at detecting medical records containing violence by employing the neural model. The training and validation of the model was performed on 76,911 records randomly sampled from the whole dataset; the test involved as many items.

The results of the categorization step are reported in Table 1: we obtained a 99% F1 score for the non-violence class. The F1 score for the V class is 86%. The neural model identified 2,291 entries as violence (V) cases; regarding the V class, the true positives amount to 2,073 out of 2,584 items, thereby yielding a .92 precision and a .80 recall. Overall, 218 false positives were detected (i.e., such data was labeled as V by the system, but annotated as NV by hospital staff).

Although we consider the obtained figures as an encouraging result, a closer inspection of the false positives revealed that in 65% of cases (that is, 141 out of 218 false positives) the system had predicted V, mistakenly annotated by the hospital staff as NV. For example, the record with text '[...] the patient reports that he had been beaten by known people, all over his body but especially on the right shoulder [...]' had been annotated as NV, while the VIDES system had predicted it as a violent case (V). In such cases the annotation is wrong. After manually correcting such errors, the precision obtained by the VIDES system raises to 97%. The updated figures are reported in Table 2. Of course, we note herein a significant +5% improvement (w.r.t. results reported in Table 1) in the precision, but also the recall and the harmonic mean benefit from this *ex post* data cleaning step.

The precision of the categorization module on the V class ensures that the explanation module (taking as input the records labeled as V at categorization time) is mostly executed on records describing injuries related to violence events. The set of records labeled by the neural model as V has then been used to assess the accuracy of the explanation module, concerned with extracting the relevant information to fill the violence frame slots.

Frame Elements Extraction Results

In order to evaluate the quality of the extracted fillers we have taken into consideration each field separately. The following metrics (standard in Information Retrieval tasks) were recorded to assess the output of the system:

- Mean Average Precision (MAP): the mean of the average precision obtained over all dataset, where the average precision is the precision of each element given as result;
- Success at 1 (S@1): the percentage of cases in which the first value was correct;
- Success at 5 (S@5): the percentage of cases in which among the first five values the correct value was returned.
- Recall at 5 (R@5): how many of the correct values were returned among the first five values.

Additionally, we developed a baseline against which we compared the output of the proposed approach. The baseline adopts the same pre-processing as the main algorithm, with the difference that it only employs semantic similarity to rank the results; the similarity threshold was also preserved, but no PoS

tag/SST filtering was employed. The detailed results are provided in Table 3.

The whole control strategy always favorably compares to the baseline, thereby showing that PoS tagging and SST information are helpful to extract the information to fill the frame slots.

Discussion

As regards as the neural network module, the VIDES system showed an optimal accuracy (.99 F1 score) in categorizing NV records, and near optimal accuracy (.88 F1 score) in categorizing V records, reporting about injuries inflicted intentionally. As regards as the latter ones, we stress the relevance of the precision (.97). The output of this module is reliable, to such an extent that it has been already used to check and to correct, although in supervised fashion, the information collected in real-word, hospital records.

The task of extracting the relevant pieces of information to fill the violence frame confirms to be a challenging and stimulating one. Different degrees of difficulty feature the recognition of the relevant frame elements. TIME and LOCATION of the violence event were individuated to a greater extent than other elements, such as the MODE-INSTRUMENT, LESION-TYPE and BODY-PART. As regards as such fields, we note that on average more information was available (e.g., MODE-INSTRUMENT was filled in 97% of cases, with 1.445 fillers, on average, over the 200 considered records), that may have been detrimental to the exact identification of such elements.

A closer inspection at the errors in the generation of the explanation may be beneficial for future improvements and for similar applications grounded on the adoption of distributed word representations paired with the filling of semantic frames. Some errors in the recognition of the AGENT originate from the fact that further persons can be mentioned in the ER report (e.g., in a sentence like 'the father reports that the patient was punched by her husband'). In such cases neither PoS nor SST information are helpful in filtering out the father as the author of the violence: this sort of errors should be dealt with through syntactic parsers (at least to individuate the dependent clause 'the patient was punched by her husband'), thus permitting to rule out 'father' as the agent.

Further errors stem from the SST filtering step: in some cases even such basic disambiguation performed through supersense tags fails, thereby undermining the filtering step. This determines a too crowded set of candidates, and these elements are not properly ranked in the subsequent stage. Errors in the SST are in principle equally distributed across all classes, but their impact is worse on frame elements having more general semantic types as admissible candidates, such as MODE-INSTRUMENT and, of course, for terms with a higher degree of polysemy. The primary role of SST information is also confirmed by the comparison between the baseline and the fully fledged VIDES system.

Many errors were caused by typos: even the trivial lack of a space between two words may prevent the tokenizer from correctly recognizing the terms involved in the linguistic expression, and tools could be adopted that have been designed for the interactive correction and semantic annotation, also with special focus on narrative clinical reports [32, 33]. Additionally, one desideratum would be individuating multiword expressions such as 'neck of the bottle' or 'lacerated bruised wound' that need to be handled as a whole (and that, conversely, cannot be dealt with in a token by token mode) [34]. Unfortunately, in the considered domain and for the considered text excerpts, standard approaches such as `mwetoolkit` [35] are frequently mislead to such an extent that their adoption does not ensure substantial processing advantage.

To improve the performance in the task of semantic frame extraction, it would be thus crucial to benefit from reliable syntactic (either dependency or constituency parsing [36, 37]) information, which unfortunately could not be attained, due to the presence of frequent ungrammatical structures and out-of-vocabulary (OOV) tokens. Also, a richer representation of the frame elements could be obtained by employing knowledge graph embedding techniques [38, 39], that can be combined with predictive models [40], although these cannot alleviate the issues stemming from the poor quality of the input. In facts, ER reports are conceived as short reports for hospital insiders, rather than as a complete, fully explicit, grammatically and syntactically correct form of communication.

This is definitely what makes them intriguing and worth research efforts, like many other forms of contemporary communication, featured by similar grammatical and syntactical traits such as, e.g., social media communication [41], and some forms of spoken language involving ill-formed spontaneous spoken language and under-specified grammars [42].

Conclusions

In this article we have investigated how to provide the categorization computed through a neural-network based classifier with explanations. In particular, we have considered the task of categorizing emergency room reports, by focusing on those containing violence events. We have illustrated the motivations underlying this kind of application: contrasting violence, by promptly tracking violent episodes as they are reported in the ER setting. On a purely scientific viewpoint, we have illustrated some of the challenges inher-

ent to performing information extraction tasks when dealing with this type of language.

The input to the VIDES system is composed by text documents that, as illustrated, can be hardly elaborated with standard (e.g., syntactic parsing) NLP techniques due to many typos, abbreviations, acronyms, and so forth. As mentioned, we have cast the present task to a particular sort of Semantic Role Labeling, where the system has to fill the slots describing a violence event, that has been previously fed to the neural model. In order to explain why a record was labeled as containing a violence-related injury, the VIDES system performs a hybrid step of information extraction by employing word embeddings, supersense tags, and PoS filtering techniques.

To the best of our knowledge, no attempt has been proposed yet to tackle this task by exploiting a synthetic (vectorial) representation for each semantic slot. Although improvements can be drawn, this approach showed to obtain encouraging results, especially for some kinds of information. It would be interesting to investigate to what extent our approach generalizes to further applications in the medical domain and to further domains, as well, since in the proposed pipeline there is no domain-specific component, thereby enabling to apply it to build explanations of different sorts of output of neural models.

Consent for publication
Not applicable.

Abbreviations
EMR: electronic medical record; ERR: emergency room report; CDC: Centers for Disease Control and Prevention; VIDES: VIOLENCE DETECTION SYSTEM; SRL: semantic role labeling; PoS: part of speech; ER: emergency room; ReLU: rectified linear unit; LSTM: long short-term memory; SST: super sense tag; SINIACA: 'Sistema Informativo Nazionale sugli Incidentiin Ambiente di Civile Abitazione', National Information System on Accidents in Civil Housing Environment; IDB: injury database; MAP: mean average precision; S@1: success in the first element of the output; S@5: success in the first five elements of the output; R@5: recall of the first five elements of the output; OOV: out of vocabulary.

Acknowledgments
We are grateful to Simone Donetti, Claudio Mattutino, and Sergio Rabellino from the Technical Staff of the Computer Science Department of the University of Turin, for their precious support with the computing infrastructures. Thanks are also due to the Competence Centre for Scientific Computing (C3S) of the University of Turin [43].

Authors' contributions
EM, DC and DPR developed the idea, implemented the method, conducted the experiments, and drafted the manuscript. MD, and CM contributed with fundamental discussions and suggestions on the overall approach; MG and AP provided critical review and the experimental data. All authors read and approved the final manuscript.

Funding
The first author was partly supported by a grant provided by Università degli Studi di Torino. This research is also supported by Fondazione CRT, RF 2019.2263.

Availability of data and materials
The data that support the findings of this study are available from the National Institute of Health (Istituto Superiore di Sanità, ISS) but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from the authors upon reasonable request, and with permission of ISS.

Ethics approval and consent to participate
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Department of Computer Science, University of Turin, Corso Svizzera 185, 10149 Turin, ITALY. [2]Servizio sovrazonale di Epidemiologia dell'ASL TO3 della Regione Piemonte, Via Sabaudia 164, 10095 Grugliasco (TO), ITALY. [3]Reparto Epidemiologia ambientale e sociale Dipartimento Ambiente e Salute (DAMSA) Istituto Superiore di Sanità, Viale Regina Elena, 299, 00161 Roma, ITALY.

References

1. Moulin, B., Irandoust, H., Bélanger, M., Desbordes, G.: Explanation and argumentation capabilities: Towards the creation of more persuasive agents. Artificial Intelligence Review **17**(3), 169–222 (2002)
2. Aamodt, A.: Explanation-driven case-based reasoning. In: European Workshop on Case-Based Reasoning, pp. 274–288 (1993). Springer
3. Roth-Berghofer, T.R.: Explanations and case-based reasoning: Foundational issues. In: European Conference on Case-Based Reasoning, pp. 389–403 (2004). Springer
4. Quinlan, J.R.: Induction of decision trees. Machine learning **1**(1), 81–106 (1986)
5. Tibshirani, R.: Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological) **58**(1), 267–288 (1996)
6. Voigt, P., Von dem Bussche, A.: The eu general data protection regulation (gdpr). A Practical Guide, 1st Ed., Cham: Springer International Publishing (2017)
7. Ras, G., van Gerven, M., Haselager, P.: In: Escalante, H.J., Escalera, S., Guyon, I., Baró, X., Güçlütürk, Y., Güçlü, U., van Gerven, M. (eds.) Explanation Methods in Deep Learning: Users, Values, Concerns and Challenges, pp. 19–36. Springer, Cham (2018). doi:10.1007/978-3-319-98131-4_2
8. Pieters, W.: Explanation and trust: what to tell the user in security and ai? Ethics and information technology **13**(1), 53–64 (2011)
9. Miller, T.: Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence **267**, 1–38 (2019)
10. Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., Müller, K.-R.: Unmasking clever hans predictors and assessing what machines really learn. Nature communications **10**(1), 1096 (2019)
11. Basile, V., Caselli, T., Radicioni, D.P.: Meaning in context: Ontologically and linguistically motivated representations of objects and events. Applied Ontology **14**(4), 335–341 (2019). doi:10.3233/AO-190221
12. Samek, W.: Explainable AI: Interpreting, Explaining and Visualizing Deep Learning vol. 11700. Springer, ??? (2019)
13. Organization, W.H.: Responding to intimate partner violence and sexual violence against women: Who clinical and policy guidelines. Technical report, World Health Organization (2013)
14. Organization, W.H., et al.: Who: addressing violence against women: key achievements and priorities. Technical report, World Health Organization (2018)
15. Leeb, R.T.: Child Maltreatment Surveillance: Uniform Definitions for Public Health and Recommended Data Elements, Version 1.0. Technical report (2008)
16. Fillmore, C.J., Baker, C.: A frames approach to semantic analysis. In: The Oxford Handbook of Linguistic Analysis, (2010)
17. Hermann, K.M., Das, D., Weston, J., Ganchev, K.: Semantic frame identification with distributed word representations. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1448–1458. Association for Computational Linguistics, Baltimore, Maryland (2014).

doi:10.3115/v1/P14-1136.
https://www.aclweb.org/anthology/P14-1136
18. Sikos, J., Padó, S.: Using embeddings to compare framenet frames across languages. COLING 2018, 91 (2018)
19. Gildea, D., Jurafsky, D.: Automatic labeling of semantic roles. Computational linguistics **28**(3), 245–288 (2002)
20. Croce, D., Giannone, C., Annesi, P., Basili, R.: Towards open-domain semantic role labeling. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pp. 237–246 (2010). Association for Computational Linguistics
21. Zapirain, B., Agirre, E., Marquez, L., Surdeanu, M.: Selectional preferences for semantic role classification. Computational Linguistics **39**(3), 631–663 (2013)
22. Roth, M., Lapata, M.: Neural semantic role labeling with dependency path embeddings. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1192–1202 (2016)
23. Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., Kuksa, P.: Natural language processing (almost) from scratch. Journal of Machine Learning Research **12**, 2493–2537 (2011)
24. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. arXiv preprint arXiv:1607.04606 (2016)
25. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
26. Miller, G.A.: WordNet: a lexical database for English. COMMUN ACM **38**(11), 39–41 (1995)
27. Ciaramita, M., Altun, Y.: Broad-coverage sense disambiguation and information extraction with a supersense sequence tagger. In: Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, pp. 594–602 (2006). Association for Computational Linguistics
28. Palmero Aprosio, A., Moretti, G.: Italy goes to Stanford: a collection of CoreNLP modules for Italian. ArXiv e-prints (2016). 1609.06204
29. Chen, D., Manning, C.: A fast and accurate dependency parser using neural networks. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 740–750 (2014)
30. Picca, D., Gliozzo, A.M., Ciaramita, M.: Supersense tagger for italian. In: LREC (2008)
31. Pitidis, A., Fondi, G., Giustini, M., Longo, E., Balducci, G., Gruppo di lavoro SINIACA-IDB, Dipartimento di Ambiente e Connessa Prevenzione Primaria, ISS: Il Sistema SINIACA-IDB per la sorveglianza degli incidenti. Notiziario dell'Istituto Superiore di Sanità **27**(2), 11–16 (2014)
32. Zvára, K., Tomecková, M., Peleška, J., Svátek, V., Zvárová, J.: Tool-supported interactive correction and semantic annotation of narrative clinical reports. Methods of information in medicine **56**(03), 217–229 (2017)
33. Wang, L., Luo, L., Wang, Y., Wampfler, J., Yang, P., Liu, H.: Natural language processing for populating lung cancer clinical research data. BMC Medical Informatics and Decision Making **19**(5), 239 (2019)
34. Constant, M., Eryiğit, G., Monti, J., Van Der Plas, L., Ramisch, C., Rosner, M., Todirascu, A.: Multiword expression processing: A survey. Computational Linguistics **43**(4), 837–892 (2017)
35. Ramisch, C., Villavicencio, A., Boitet, C.: Mwetoolkit: a framework for multiword expression identification. In: LREC, vol. 10, pp. 662–669 (2010). Valletta
36. Ivanova, A., Oepen, S., Øvrelid, L.: Survey on parsing three dependency representations for english. In: 51st Annual Meeting of the Association for Computational Linguistics Proceedings of the Student Research Workshop, pp. 31–37 (2013)
37. De Mori, R.: Spoken language understanding: a survey. In: 2007 IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU), pp. 365–376 (2007). IEEE
38. Wang, Z., Zhang, J., Feng, J., Chen, Z.: Knowledge graph embedding by translating on hyperplanes. In: Twenty-Eighth AAAI Conference on Artificial Intelligence (2014)
39. Goyal, P., Ferrara, E.: Graph embedding techniques, applications, and performance: A survey. Knowledge-Based Systems **151**, 78–94 (2018)
40. Ma, F., Wang, Y., Xiao, H., Yuan, Y., Chitta, R., Zhou, J., Gao, J.: Incorporating medical code descriptions for diagnosis prediction in healthcare. BMC Medical Informatics and Decision Making **19**(6), 1–13 (2019)
41. Danescu-Niculescu-Mizil, C., Gamon, M., Dumais, S.: Mark my words!: linguistic style accommodation in social media. In: Proceedings of the 20th International Conference on World Wide Web, pp. 745–754 (2011). ACM
42. Wang, Y.-Y.: A robust parser for spoken language understanding. In: Sixth European Conference on Speech Communication and Technology (1999)
43. Aldinucci, M., Bagnasco, S., Lusso, S., Pasteris, P., Rabellino, S., Vallero, S.: OCCAM: a flexible, multi-purpose and extendable HPC cluster. Journal of Physics: Conference Series **898**(8), 082039 (2017)

Figure Legends

Figure 1 Example of frame extraction for a sentence. A short description of the figure content should go here.

Figure 2 The neural network architecture. The neural architecture employed for the categorization task.

Tables

Table 1 Precision, Recall and F1 scores for violence (V) and non-violence (NV) classes on the test set.

Class	P	R	F1
NV	.99	1.0	.99
V	.92	.80	.86

Table 2 Precision, Recall and F1 scores for violence (V) and non-violence (NV) classes on the test set, after correction of the mistakenly annotated false positives.

Class	P	R	F1
NV	.99	1.0	.99
V	.97	.81	.88

Table 3 Results for the explanation algorithm along with the baseline.

Run	Field	MAP	S@1	S@5	R@5
Baseline	Agent	.12	.12	.14	.13
	Mode-Instrument	.24	.22	.29	.27
	Time	.73	.74	.74	.73
	Location	.50	.51	.51	.50
	Body-Part	.15	.18	.26	.18
	Lesion-Type	.30	.36	.37	.31
Main algorithm	Agent	.58	.59	.60	.58
	Mode-Instrument	.28	.31	.32	.29
	Time	.80	.82	.82	.80
	Location	.90	.90	.90	.90
	Body-Part	.49	.57	.57	.49
	Lesion-Type	.40	.45	.47	.41

Figures

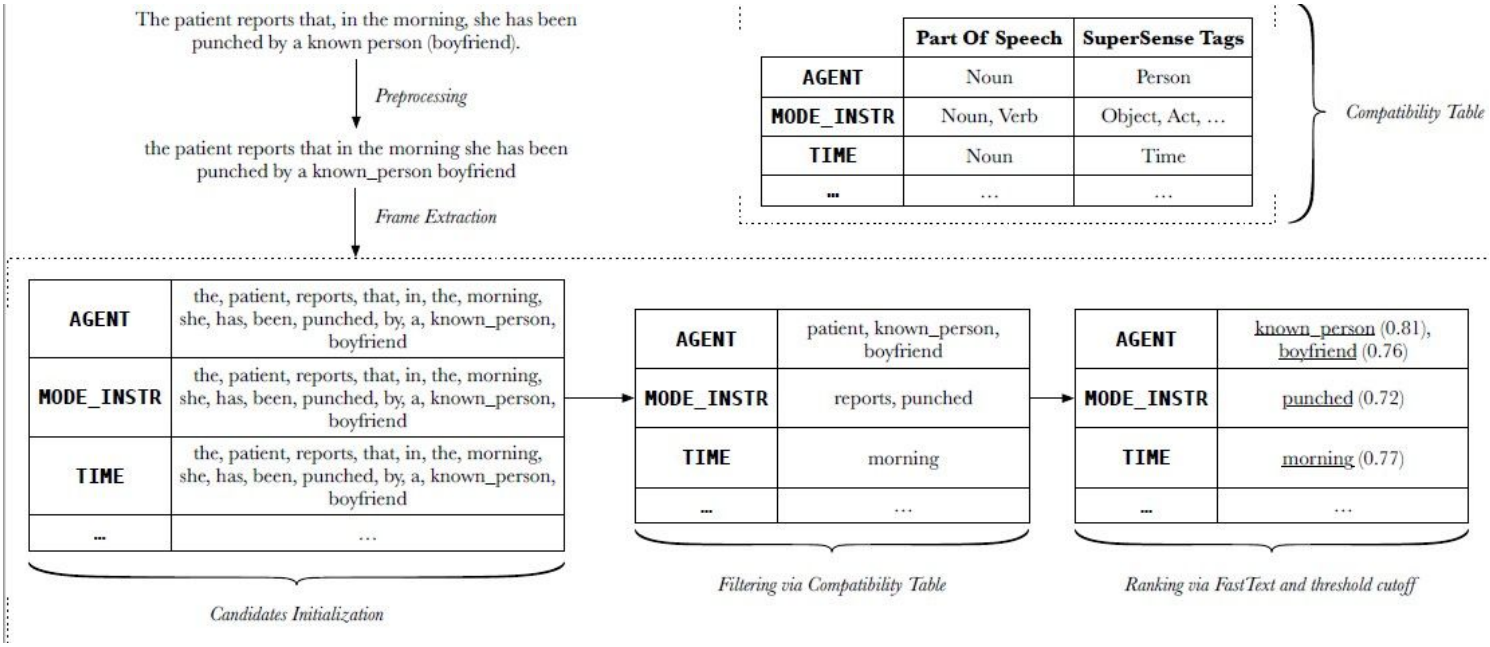

Figure 1

Example of frame extraction for a sentence.

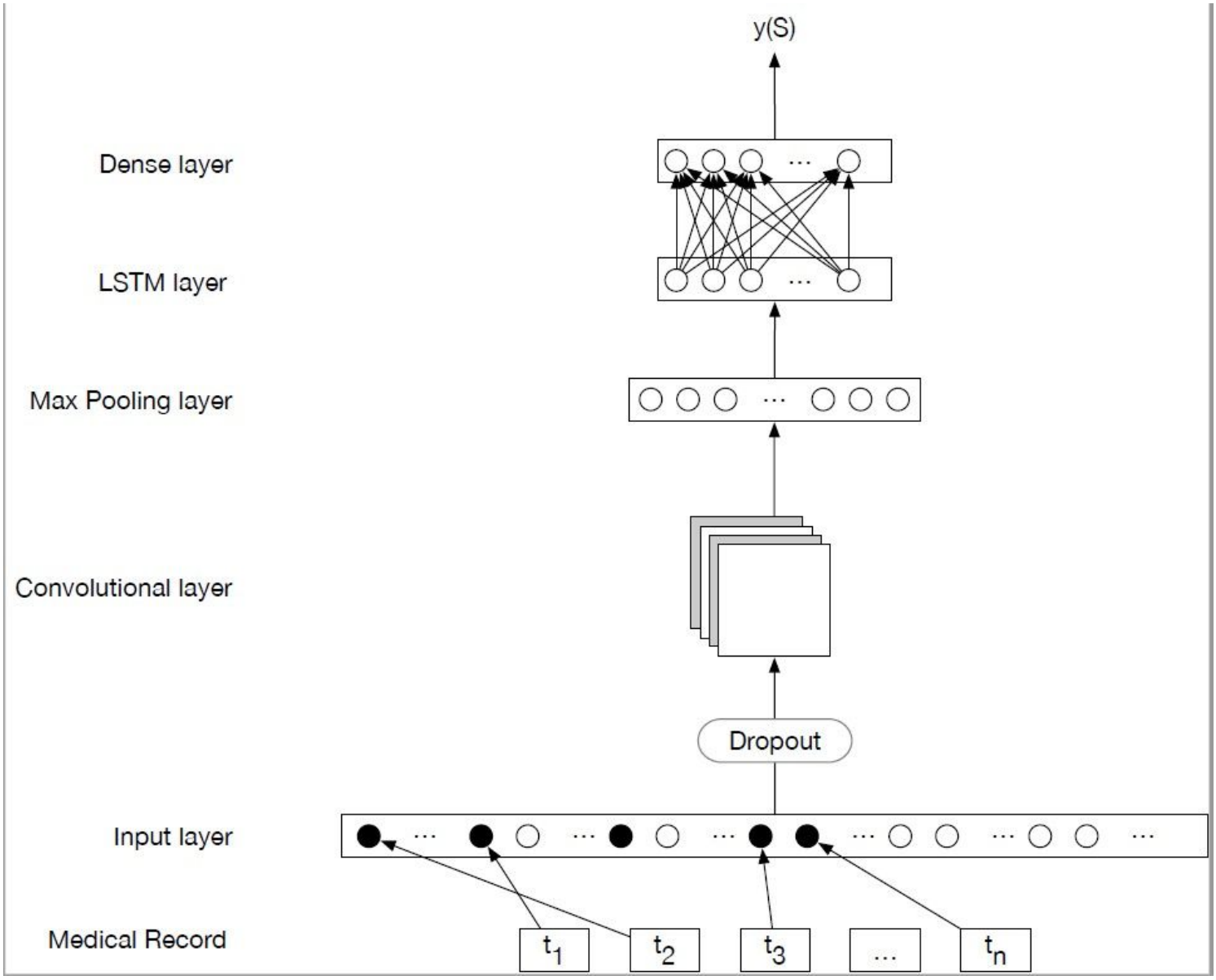

Figure 2

The neural network architecture. The neural architecture employed for the categorization task.