Explainable Sentiment Analysis with Applications in Medicine

Chiara Zucco*, Huizhi Liang†, Giuseppe Di Fatta†, Mario Cannataro*
*Data Analytics Research Center
Department of Medical and Surgical Sciences
University “Magna Græcia”
Viale Europa, 88100, Catanzaro, Italy
chiara.zucco@studenti.unicz.it, cannataro@unicz.it

†University of Reading
Department of Computer Science
University of Reading
Reading, RG6 6AY, UK
{huizhi.liang, g.difatta}@reading.ac.uk

***Abstract*—Sentiment Analysis can help to extract knowledge related to opinions and emotions from user generated text information. It can be applied in medical field for patients monitoring purposes. With the availability of large datasets, deep learning algorithms have become a state of the art also for sentiment analysis. However, deep models have the drawback of not being non human-interpretable, raising various problems related to model's interpretability. Very few work have been proposed to build models that explain their decision making process and actions. In this work, we review the current sentiment analysis approaches and existing explainable systems. Moreover, we present a critical review of explainable sentiment analysis models and discussed the insight of applying explainable sentiment analysis in the medical field.**

***Index Terms*—explainable models; AI; Machine Learning; Deep Learning**

I. Introduction

Capturing the opinions or feelings of others has always been a central point in verbal or written communication. This has led to the emergence of Sentiment Analysis, the subarea of Affective Computing grouping a series of Machine Learning and NLP methods, techniques, and tools with the aim of detecting and extracting “sentiment” from written text. Extracting knowledge related to opinions and emotions from this large amount of unstructured texts, has many practical applications ranging from commercial applications such as marketing and financial analysis, product reviews, brand monitoring, customer services, recommendation systems, to applications in healthcare field.

Current sentiment analysis techniques, can be grouped into three main approaches: lexicon based, Machine Learning and hybrid approaches. More dominant approaches are mainly linked to Machine Learning techniques and, in particular, Deep Neural Networks (DNNs) model, as for example Convolutional Neural Networks (CNNs, Recurrent Neural Networks (RNNs), and also Deep Belief Network (DBNs) [1]. They have shown a lot of success in many Sentiment Analysis tasks and in other fields such as computer vision [2].

However, it is common knowledge that current Machine Learning techniques suffer from three major problems that are particularly limiting in the context of Natural Language Processing (NLP), namely: generalization problems, coherence issues and lack of transparency [3]. These three points, are related to a topic that is currently generating particular interest in the scientific community: the interpretability of AI systems.

For example, the non-linear and nested structure of DNNs that improves accuracy in prediction models, does not allow to understand how, starting from a human-understandable input such as an image or a piece of text, the flow of information reaches a predicted output. In this sense, DNNs are regarded as black-boxes. Moreover, since DNNs are generally trained on large amount of data produced as digital traces of real human activities that may contain prejudice or biased: this may lead to predictions that can be in some way discriminative. An example can be found in [4], in which by analyzing 8000 cases using COMPAS, a criminal risk assessment tool, authors found racial disparities in risk scores, even though data did not include sensitive information as race.

Another emblematic example concerns the application of Machine Learning (ML) in the health field and in particular the prediction of the probability of death in patients with pneumonia, to optimize costs by managing patients at low risk as outpatients and admitting high risk patients [5]. Although the neural network model resulted as the most accurate, it was preferred to use a logistic regression model, because it was considered too risky to adopt a model that would not allow medical professionals to understand why and how the decision was made. In fact, a rule-based model extracted the rule that the mortality risk of patients having a history asthma, was lower compared to patients having a different medical history. This rule reflected a real pattern, in the sense that asthmatic patients with pneumonia receive a more

978-1-5386-5488-0/18/$31.00 ©2018 IEEE

aggressive care compared to other patients and this common-sense procedure was so effective to lower the risk of death in patients with asthma history.

Moreover, the need of explanatory systems is required by impending regulations like the General Data Protection Regulation (GDPR), recently adopted by the European Union. The GDPR regulates the collection, storage and use of personal information, stating that *the data controller shall implement suitable measures to safeguard the data subject's rights and freedoms and legitimate interests, at least the right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision*[...] *Decisions* [. . .] *shall not be based on special categories of personal data unless suitable measures to safeguard the data subject's rights and freedoms and legitimate interests are in place.* So, in order to increment trust in black-box models, the concept of accuracy alone cannot be sufficient but it have to be integrated to the request that the model can be also "explainable".

Currently, there are some work proposed explainable models. For example, Local Interpretable Model-agnostic Explanations (LIME) [6], [7], Layer-Wise Relevance Propagation [8], DeepLift [9], Contextual Explanation Networks (CEN) [10]. Very few works discuss explainable Sentiment Analysis models [11],[6]. How to develop explainable sentiment analysis models and increase the explainability of decision making of sentiment analysis models still need to explored.

Aim of this work is to discuss explainable models, and giving some insight in applying them when developing Sentiment Analysis applications in medicine. The rest of the paper is organized as follows. In Section II main Sentiment Analysis tasks and methodologies are presented. In Section III general explanation models are discussed, while in Section IV some insight in developing explainable SA systems for medical application are discussed. Finally, Section V concludes the paper.

II. Existing Sentiment Analysis models

In common language, the word sentiment has at least a double meaning: sentiment seen as sensation or emotion, and sentiment as a personal thought or opinion. If sentiment thought as emotion can be extracted from gestures, expressions, tone of voice and written or verbal contents, sentiment seen as opinion is much more linked to verbal or written content and, therefore, to text-based input. That is why Sentiment Analysis of a generic "text unit" encompasses problems related to the recognition of both emotions and opinions expressed in that unit [12].

On the basis of the analyzed text unit, Sentiment Analysis has been performed at three levels: document level, sentence level and aspect or, more generally, sub-sentence level.

In terms of tasks, sentiment analysis can be considered, as a big suitcase of NLP problems [3]. However, the most popular task is "polarity recognition" that can be synthesized as the problem of determining whether a text unit contains a positive/negative or neutral opinion. Polarity classification can be also performed at a finer-grained level, trying to extract from a text unit some opinion strength levels, expressed as a score taking values into a real range, or as a rank scale (such as Very positive, Positive, Neutral, Negative, Very negative).

Recently, thanks to the growing interest into social monitoring and its applications in social sciences and in medical and psychological fields [13], [14], [15], another Sentiment Analysis task related to "emotion classification" has gained attention. The aim of the emotion recognition task is to identify which emotions of a list of basic emotions, provided by cognitive and psychological models, is expressed in a text unit.

TABLE I
Emotions definitions.

Plutchik [16]	Acceptance, anger, anticipation, disgust, joy, fear, sadness, surprise
Arnold [17]	Anger, aversion, courage, dejection, desire, despair, fear, hate, hope, love, sadness
Ekman [18]	Anger, disgust, fear, joy, sadness, surprise

With reference to methodologies, three main approaches can be identified: lexicon-based, Machine Learning, and hybrid approaches [19], [20].

In **Lexicon-based approaches** the overall sentiment is then extracted on a basis of a defined set of rules. These rules generally relies on a lexicon, i.e. a set of tuples containing a word along with that word's sentiment orientation (both in terms of polarity or emotions), created in a manual (dictionary based) or automatic (corpus-based) fashion. There are some downside of lexicon-based approaches: one is that people express sentiments in different ways and using different words, so lexicons are required to evolve in very short times. Another issue is that words polarity relies on the context (for a car to be fast is good, while maybe no one wants a fast novel) so there is the need of huge and suitable lexicons. A third point is that lexicon-based approaches do not take into account the semantic structure of the text, since they solely rely on the occurrences of words that explicitly represent a sentiment. A third disadvantage is that lexicon-based systems get very complex quickly.

In a general **Machine Learning approach**, after being preprocessed, the text unit is vectorized in a suitable representation with classical bag-of-words (or bag-of-ngrams) approaches or relying on more recent word embedding systems [21]. From an higher-dimensional and sparse vector space a word embedding learns a new representation of text into a dense lower-dimensional vector. Although most common used word embedding systems are Word2Vec and Glove, also attempts to propose a Sentiment-specific word embedding have been made [22], [23]. In terms of approaches, Machine Learning techniques for Sentiment Analysis encompass supervised and unsupervised learning methods. The supervised methods are the dominant approaches because a Sentiment Analysis task is usually modeled as a classification or a regression problem. In terms of used algorithms, Machine Learning techniques for Sentiment Analysis can be categorized in: i) shallow approaches and ii) deep-learning approaches.

Shallow Machine Learning approaches were mainly used in the early development of Sentiment Analysis. The work in [24] was one of the first to apply Machine Learning classification algorithms in order to determine binary polarity detection in a movie review dataset. They applied Naïve Bayes, Maximum Entropy and Support Vector Machines (SVM) and the latter gave the best results. SVM is also used as a classifier in emotion recognition toolkit presented in [25]. The toolkit can used be for detecting emotions from text and it also can be fed with manually annotated data in order to train a classifier on a customized emotion theory. However, the classifier performances was not compared with some baseline. Ensemble methods techniques such as Bagging and Boosting were also used in order to enhance base learners performance, both for polarity and emotion classification tasks [26], [27].

The increasing interest in sentiment analysis is also linked to the possibility of extracting latent or explicit emotions or opinions from social media posts. Data coming from social media are unstructured and in general they are stored in large-scale databases. In order to perform sentiment analysis with a massive amount of data, as for example social media data, **deep learning algorithms** have recently gained a lot a popularity, by achieving state-of-the-art results in various tasks [1].

Roughly speaking, Deep Neural Networks (DNNs) are a set of algorithms designed for modeling a task by using an Artificial Neural Network composed of multiple layers, i.e. multiple levels of non-linear and nested operations. The DNN's capability of modeling non-linear tasks has made them extremely popular in both research and application fields in relation to a wide range of tasks, and especially for predictions problems.

An approach for modeling higher level concepts in text, is to represent a text unit as a sequence of words and then perform learning in terms of spatial patterns. Several DNN's architectures have been proposed in Sentiment Analysis, taking into account Autoencorder Neural Networks [28], Convolutional Neural Networks (CNNs) [29], [30], Recurrent Neural Networks (RNNs) with or without attention mechanism, and also Deep Belief Network (DBN) [31], [32].

A special type of RNN is Long Short Term Memory (LSTM) network in which the network architecture can learn long-term dependencies by updating hidden state representation in a suitable way [33]. This is of particular importance in text categorization because LSTM can learn a semantic representation of text, but it also is considered a state-of-the art algorithm for performing prediction on time series features that can be of primary importance in monitoring systems. In fact, in [34], LSTM was used to predict depressed mood in self-reported histories.

A major exponential drawback that a common Deep Learning approach may lead to, is the difficulty in generalizing predictions to novel viewpoints. This is because modeling new viewpoints implies affine transformations. When dealing with new affine dimensions, an exponential increase of training set size or in the size of translated replicas of learned feature detectors is required.

More recently, Capsule Networks [35] and in particular a variant of Capsule Networks that takes into account an iterative routing process between layers, called Dynamic routing [36], have recently gained a lot of interest. Capsule Networks have shown to potentially address the aforementioned generalization issue. Performance of Capsule Networks with dynamic routing was also been investigated for text classification [37], showing interesting results, especially in addressing multi-labeled problems. However, for what concern Sentiment Analysis tasks, even if some efforts were made in this direction [38], [37], a more in-depth evaluation has still to be made.

In order to find a trade-off between the Machine Learning drawback related to the dataset size required for learning and the lexicon based problems discussed above, hybrid approaches that combine Machine Learning and lexicon-based approaches, were also taken into account [39][40].

From this brief overview, it is quite evident that Machine Learning and, in particular, deep learning algorithms, have become a dominant approach also in text classification field, reaching an excellent accuracy. However, as stated in the Introduction, most of the aforementioned approaches are used as a "black-box". In order to develop systems to address real world problems, with a special mention in the medical field, a discussion on how predictions can be made more explainable still needs to be made.

III. Explainable models

Even though also in sentiment classification, many Machine Learning approaches and, in particular, deep learning models are now considered state-of-the art approaches, their "black-boxes" nature is leading to an ever increasing interest toward "Exaplinable AI", and "Explainable Machine Learning". There is not a uniform definition of explainable or interpretable Machine Learning, nor a clear difference or a recognized equivalence between "Explainable" or "Interpretable" Machine Learning: in some cases they are used as synonyms, while in others a distinction is made.

Although investigating the various interpretations of explainable and interpretable models is beyond the scope of this paper, it is considered appropriate to present the theoretical framework of interpretability and explainability. For this reason, instead of following some specific approach, in this section a brief overview of different definitions of "explainable/interpretable" ML is given, the desiderata of explainable systems will be discussed as well as some proposed taxonomy of existing approaches.

A. Basic definitions

Generally speaking, even though the verbs "to interpret" and "to explain" both imply making something clear or understandable, if the former is more related to the idea of an interpreter translating from one language to another one, that is more comprehensible to the audience, the latter also implies to make something intelligible when is not immediately obvious or entirely known[1].

In [41], interpretation is defined as *the process of mapping an abstract concept (e.g. a predicted class) into a domain that the human can make sense of,* while an explanation is *the collection of features of the interpretable domain, that have contributed for a given example to produce a decision (e.g. classification or regression)*.

In [42] interpretability in Machine Learning is defined as *the ability to explain or to present in understandable terms to a human*. This definition is also reported in the survey of Guidotti et al. [43], in which an explanation is thought as *an "interface" between humans and a decision maker that is at the same time both an accurate proxy of the decision maker and comprehensible to humans*.

Starting from the consideration that existing works in explainable artificial intelligence only take into account researchers' idea of what are good explanations, the extensive work of Miller [44], gives a definition of explainability by taking into account insights from social sciences. His major findings are that human explanations tend to be: contrastive, selected among others, that a most-likely explanation does not always implies to be the best possible explanation, and that explanations are social, in the sense that they require a transfer of knowledge. Despite the lack of a shared definition, a more shared line is found in delineating interpretable models desiderata: interpretability (intended as human-understandability), coherence, generalizability and explanatory power [45].

In the state of the art, there is a small set of recognized interpretable models such as decision tree or gradient boosted trees, rules-based and linear models. A simpler way to evaluate interpretability is via a quantifiable proxy [42]: the interpretable algorithm is presented as an optimization of a class models that already are claimed to be interpretable [46].

A more general evaluation of the goodness of an interpretable system can rely on the evaluation of the four characteristics mentioned above. However, for what concerns the formalization of suitable measures and methods for evaluating explainable systems, work in this field is still at an early stage and we believe that further research efforts are needed in this direction.

In general, coherence and fidelity have been assessed through accuracy measures. In particular, if considering the outcome of the original model as an oracle, an external interpreter model has high levels of fidelity when it reaches high levels of accuracy.

As mentioned in the Introduction, in systems where the interpretation is provided by an external model with respect to the original model, the interpretability is closely linked to the idea of providing a representation of the original model, especially through the visualization of the most relevant. In this type of approach, the quality of the produced explanation is deduced through strategies that evaluate the visualization of what the system has learned. An approach of this kind is presented in the work of Samek et al. [47], in which a qualitative measure is proposed to evaluate ordered collections of pixels such as heatmaps, by region perturbation. This technique was also used in the context of text classification [11]. For assessing interpretability, to the best of our knowledge, the most used approaches rely on human evaluation and, in particular, on expert evaluation [48], [49], [50].

B. External explainer models

Also with respect to methods, there is not a clear categorization. The taxonomy presented in [45] grouped main approaches into three categories: rule-extraction, attribution and intrinsic methods.

Rule-extraction methods aim at extracting logical rules, for example IF-THEN, AND/OR or M-of-N rules and also decision trees by using the same input on which the DNN is trained and comparing the output of the trained DNN model. In the review of Hailesilassie [51] on Rule Extraction Algorithm for DNNs, three main approaches are highlighted, i.e. Decompositional, Pedagogical, and Eclectic. Rule-extraction explaination can be found in the apporaches of [52], [53], [54].

Attribution or relevance methods are actually the most popular methods and aim to explain the model in terms of how relevant is a feature for the final prediction. Attribution methods are often visualized in term of heatmap and performance evaluations usually are related to perturbation, ablation and influence studies. Perturbation strategies are common in so called Model Agnostic methods, most popular of which is Local Interpretable Model-agnostic Explanations (LIME) [6], [7]. Other popular approaches for explaining deep networks prediction are Layer-Wise Relevance Propagation [8] and DeepLift [9]. There are some conflicting views regarding the lackness in sensitivity of local explanation methods [55], [56].

Intrinsic methods the improvement in the interpretability of a model is not done in a *post-hoc* fashion but by enhancing internal interpretable representations with methods that are part of the DNN architecture, as for example by providing an interpretable loss function. With respect to relevance methods, intrinsic methods have the advantage of increasing fidelity and decreasing the complexity of the resulting explanation.

One of the major issues of current approaches for interpretability is that they address the problem from a technical standpoint, providing explanations that are not meant for the end user [44]. For example, when seeing an image, people recognize a cat, they do not justify the answer by selecting the pixels on which they are based, but perhaps in terms of "ears", nose, tail etc. Intuitively, this is one reason why the Capsule Networks have shown potential intrinsic explainability properties, because they are constructing a part-whole relationship that can be seen as a relevance path [57]. Since the ultimate goal of an explicable system is to provide explanations to people, then it is necessary to take into account also the domain and the levels of knowledge of the people to whom the explanation is to be provided. Interesting approaches in this

[1] https://www.merriam-webster.com/dictionary/interpret

sense can be found in a few works related to “self-explainable” models.

C. Self-explanatory models

As emphasized in the recent paper of Doran et. al [58], most of the contributes in explainable AI field are focused on enabling explanations of models decisions, while few attempts are made to build models that “generate” explanations. With the expression “self-explanatory” we refer to a model that in addiction to some other task, also generates a line of reasoning in order to explain to the user which decision-making process was followed in terms of input characteristics and by using to high level concepts that are user-understandable. Since understanding decision-making process that leads from the input to that specific class, should be the final goal of explainable systems, this section is devoted to present two different examples of “generating explanation” approaches.

In [48], Hendricks et al. propose a captioning systems providing a description that also includes explanations of why the predicted category is the most suitable for that specific input image. Their approach consists of a long term recurrent convolutional network [59] consisting of a convolutional neural network which extracts highly discriminative image features with the advantage of a compact bilinear classifier, and two stacked Long-Short Term Memory (LSTM) networks learning how to generate a description conditioned on the features extracted by the CNN modules. During training, the model receives an instance consisting of an image, a category label and a ground truth sentence and is trained to predict each word in a ground truth sentence, by minimizing a relevance loss function and producing captions describing the input image. In order to provide a sentence generation that is also discriminative on the basis of visual extracted features, another discriminative loss is defined by using a reinforcement learning approach. An aspect which should be highlighted is that for evaluation authors used both automatic metrics and human evaluations. In particular, automatic metrics were introduced to assess image and class relevance measure for the produced explanation. However, in order to assess if the model provides “explainable justification”, a human evaluation was needed.

In [50], a Reinforcement Learning method in which the agent generates explanations for justifying its actions and its strategies is presented. Reinforcement Learning is a techniques in which an “agent” learn how to attain a goal in response to the changing state of an “environment”, and by improving its actions with experience. Reinforcement learning (RL) encompasses the fields of dynamic programming and supervised learning in order to provide powerful machine-learning systems, and the recent approaches of incorporating deep learning architectures in a RL framework, has led to powerful techniques successfully applied in different areas [60]. As for Deep Learning, also RL techniques lead to opaque systems, that are not able to explain why the agents take the route leading to the final goal.

In the work of van der Waa et al. [50], the agent answers contrastive questions about its actions on the basis of expected consequences of its policy. In particular, since in contrastive questions facts are posed against counterfactual cases, or foil, in a “why you did X instead of doing Y?” where X is the fact and Y is the foil. Considering as a fact an entire learned policy, a “foil policy” is obtained on the basis of the foil in the user’s question.

By learning a state transition model that samples the effect of both the two policies, expected consequences are modeled as a Markov Chain of states visits under both the two policies. The generated explanation, given in terms of state-actions and rewards, are translated into a more descriptive state classes and outcomes by following the approaches of [61] and [62]. Also in this paper, in order to assess some model evaluation, human evaluation was considered.

Most of the explainable methods presented in this Section have been proposed and tested mainly to explain models related to image classification problems. Some of the models previously discussed have also been applied or extended to text classification and to Sentiment Analysis. We will discuss these models in the next section.

IV. Explainable models in Sentiment Analysis with application in medicine

Text classification approaches can be directly applied in Sentiment Analysis, so in reviewing explainable models in Sentiment Analysis, also more general approaches for text classification will be taken into account. As stated in the previous section, a basic approach for assessing interpretability is by recurring to a quantifiable proxy. In sentiment analysis lexicon-based approaches are considered white boxes additive models and therefore they are trivially interpretable. Starting from these considerations, the work of Clos et al. [63] proposes a hybrid classification model that performs lexicon based classification by using a lexicon generated with a learning procedure. Another form of interpretability can be trivially assessed in the work of [64] in which at each node of a parsing tree sentiment polarity can be classified [11]. A recent work of Shahroudnejad et al. [57] potentials intrinsic explainability properties Capsule Networks were discussed. Furthermore author showed the possibility of transforming deep learning architectures in to transparent networks via incorporation of capsules in different layers. However, this discussion did not take into account explainability improvement of Capsule Networks for text classification problems. We think that this is a point that needs to be further investigated.

Referring back to the taxonomy presented in the previous section, several attribution approaches have been applied or adapted for explaining text classification models. For example, LRP model [8] was also applied in order to extract from text which words were most relevant for a CNN’s classifier trained on a topic categorization task [65]. In [65], LRP was applied in order to explain CNNs models for generic NLP classification tasks, then an extension of LRP model was presented providing an explainable model of recurrent networks architectures for Sentiment Analysis classification tasks. In particular, this model was applied to bi-directional

LSTM model trained on the Stanford Sentiment Treebank dataset for a five-class polarity classification task. This approach can actually enhance trust in the model, by showing that words with highest relevance are indeed words with strong semantic meaning. However, following the consideration that good explanations are contrastive [44], we claim that the LRP model alone does not succeed in making clear the original model's underlying decision process because, for example, it's does not allow to understand why the original model assigns a certain polarity instead of another to a given text unit.

Another attribution approach that was discussed for Sentiment Analysis, is LIME [7]. Systems evaluation were performed through different human-based assessments. A good critical point was highlighted in the work of Al-Shedivat et al., that compared LIME with their proposed "self-explainable" approach Contextual Explainable Networks (CEN) [10] on MNIST and IMDB image datasets. Results showed that when injecting different noise levels in the features, LIME continued to approximate predictions well, while CEN performance was negatively affected from low quality representation. Results of this evaluation substantially showed that LIME model can lead to misleading explanations. This is another main drawback of "external explanation" models, that can encourage to further investigate into "self-explainable" models.

To the best of our knowledge, only few works have considered to produce a textual classifier that can also generate explanations. In [66], authors proposed an Explicit Factor Model (EFM) based on phrase-level sentiment analysis, in order to generate explainable recommendations. As in many previously cited works, also in this case explainability was assessed by showing that generated recommendations were more influential on user's purchasing behavior.

In [67], authors propose a modular neural framework to automatically generate interpretable and extractive rationale. Extractive rationale is intended as the concise and sufficient subsets of words extracted from the input text. This subset is requested to be interpretable and to represent a suitable approximation of the original output, in the sense that a rationale has to provide nearly the same prediction (target vector) as the original input. A major contribution of this work is that extractive rationale is considered as a latent variable, generated in an unsupervised fashion and that the generation is incorporated as a part of the overall learning process leading to a justified neural network prediction. A minor point is that an interpretable extractive rationale is defined as a coherent and concise subset of text, but interpretability is solely assessed as a consequence of the extraction process.

Consistency and interpretability are not quantitatively evaluated and, under the assumption that better rationales achieve higher performance, evaluations of the proposed framework are based on accuracy measures such as mean average precision (MAP) for a similar text retrieval task and Mean Squared Error on the test set for the multi-aspect sentiment prediction task.

A. Insight in developing explainable Sentiment Analysis applications in Healthcare

There are several works related to Sentiment Analysis for the detection, prediction and monitoring of healthcare related problems. Even if studies performing Sentiment Analysis on clinical text exists in literature, the majority of SA applications in healthcare field mainly rely on the analysis of text extracted from Twitter. In particular, there are studies analyzing tobacco and ecigarette-related and drug abuse–related tweets [68], [69] or identifying adverse effects in medical use of pharmaceutical drugs [70]. In psychology, Sentiment Analysis applications have been proposed to: assess concern levels in suicide-related tweets, using both human coders and an automatic machine classifier [71]; to detect depression symptom in tweets [72],and also for the early detection of a whole range of depression-related disorders, such as post traumatic stress [73], bipolar disorder [74], [75], or seasonal affective disorder [76] and [77].

In order to develop a Clinical decision support systems (CDSSs) clinicians can trust, the system has to be good enough to ensure that i) clinicians understand the predictions (in the sense that predictions have to be consistent with medical knowledge), ii) decision will not negatively affect the patient, iii) the decisions are ethical, iv) the system is optimized on complete objectives, and v) that the system is accurate and sensible patient data are protected. Each of the previous requirements is related to one among five scenarios identified in the work of Doshi-Valez and Kim [42], for which it is necessary to request interpretability of the used model [42], [78].

The necessity to build explainable AI systems for the medical domain is well explained in the work of Holzinger et al. [78], in which authors discussed the insight in developing explainable AI systems for supporting medical decisions. The focus was posed on three data sources: images, *omics data and text. In this work, authors assessed the need of new strategies for presenting human-understandable explanations that can also take into account sentiment analysis methodologies. Moreover, a possible way to develop explainable AI system for healthcare domain was proposed. The proposal combines: i) a Human-in-the-Loop Machine Learning approach, where the model is built continuously and improves over time via feedback and interaction of domain experts and, ii) an interpretable disambiguation system that can provide common sense and human-readable reasons why in a given context, a given sense was detected [79].

Questioning whether this proposed framework could effectively lead to a good explanatory system and if and how supplementing it with explainable Sentiment Analysis methodologies may bring a significant contribution, constitutes a key point for further studies.

V. Conclusion

This paper highlighted the need for an explainable sentiment analysis in order to extract human-reliable knowledge related to opinions and emotions from user generated text information, especially for medical applications.

Since current state-of-the-art Sentiment Analysis techniques mainly rely on deep neural networks, they generate highly non human-interpretable models. This raises problems of trust, biases, privacy issue, etc. To address this problem, explainable models have recently gained attention. However, very few work have been proposed to build models that really explain their decision making process and actions.

After having introduced most common methodologies of Sentiment Analysis, the main contribution of the paper was: to point out more relevant methodologies, to identify the gap in applying current explainable approaches to Sentiment Analysis models, and outlining future directions that can not only lead to an explainable Sentiment Analysis system, but can also give a significant contribution towards the development of an explainable Clinical Decision Support System.

REFERENCES

[1] L. Zhang, S. Wang, and B. Liu, "Deep learning for sentiment analysis: A survey," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, p. e1253, 2018.

[2] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, p. 436, 2015.

[3] E. Cambria, S. Poria, A. Gelbukh, and M. Thelwall, "Sentiment analysis is a big suitcase," *IEEE Intelligent Systems*, vol. 32, no. 6, pp. 74–80, 2017.

[4] J. Angwin, J. Larson, S. Mattu, and L. Kirchner, "Machine bias: There's software used across the country to predict future criminals," *And it's biased against blacks. ProPublica*, 2016.

[5] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad, "Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2015, pp. 1721–1730.

[6] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should i trust you?: Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. ACM, 2016, pp. 1135–1144.

[7] ——, "Model-agnostic interpretability of machine learning," *arXiv preprint arXiv:1606.05386*, 2016.

[8] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PloS one*, vol. 10, no. 7, p. e0130140, 2015.

[9] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," *arXiv preprint arXiv:1704.02685*, 2017.

[10] M. Al-Shedivat, A. Dubey, and E. P. Xing, "Contextual explanation networks," *arXiv preprint arXiv:1705.10301*, 2017.

[11] L. Arras, G. Montavon, K.-R. Müller, and W. Samek, "Explaining recurrent neural network predictions in sentiment analysis," *arXiv preprint arXiv:1706.07206*, 2017.

[12] A. Yadollahi, A. G. Shahraki, and O. R. Zaiane, "Current state of text sentiment analysis from opinion to emotion mining," *ACM Computing Surveys (CSUR)*, vol. 50, no. 2, p. 25, 2017.

[13] F. Ciullo, C. Zucco, B. Calabrese, G. Agapito, P. H. Guzzi, and M. Cannataro, "Computational challenges for sentiment analysis in life sciences," in *2016 International Conference on High Performance Computing & Simulation (HPCS)*. IEEE, 2016, pp. 419–426.

[14] C. Zucco, B. Calabrese, and M. Cannataro, "Sentiment analysis and affective computing for depression monitoring," in *Bioinformatics and Biomedicine (BIBM), 2017 IEEE International Conference on*. IEEE, 2017, pp. 1988–1995.

[15] B. Calabrese, M. Cannataro, and N. Ielpo, "Using social networks data for behavior and sentiment analysis," in *Proceedings of the 8th International Conference on Internet and Distributed Computing Systems, IDCS 2015*.

[16] R. Plutchik, *Emotion. Theory, Research and Experiences*, P. R and K. H, Eds. Academic Press, 1980.

[17] M. Arnold, *Emotion and Personality*. Columbia University Press, 1960.

[18] P. Ekman and V. Wallace, *Unmasking the face*. Malor Book, 2003.

[19] W. Medhat, A. Hassan, and H. Korashy, "Sentiment analysis algorithms and applications: A survey," *Ain Shams Engineering Journal*, vol. 5, no. 4, pp. 1093–1113, 2014.

[20] K. Ravi and V. Ravi, "A survey on opinion mining and sentiment analysis: tasks, approaches and applications," *Knowledge-Based Systems*, vol. 89, pp. 14–46, 2015.

[21] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural language processing (almost) from scratch," *Journal of Machine Learning Research*, vol. 12, no. Aug, pp. 2493–2537, 2011.

[22] D. Tang, F. Wei, N. Yang, M. Zhou, T. Liu, and B. Qin, "Learning sentiment-specific word embedding for twitter sentiment classification," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, vol. 1, 2014, pp. 1555–1565.

[23] L.-C. Yu, J. Wang, K. R. Lai, and X. Zhang, "Refining word embeddings for sentiment analysis," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 534–539.

[24] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? sentiment classification using machine learning techniques," in *EMNLP '02 Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language PRocessing*, 2002, pp. 79–86.

[25] F. Calefato, F. Lanubile, and N. Novielli, "Emotxt: a toolkit for emotion recognition from text," *arXiv preprint arXiv:1708.03892*, 2017.

[26] G. Wang, J. Sun, J. Ma, K. Xu, and J. Gu, "Sentiment classification: The contribution of ensemble learning," *Decision support systems*, vol. 57, pp. 77–93, 2014.

[27] N. Gupta, M. Gilbert, and G. D. Fabbrizio, "Emotion detection in email customer care," *Computational Intelligence*, vol. 29, no. 3, pp. 489–505, 2013.

[28] H. Sagha, N. Cummins, and B. Schuller, "Stacked denoising autoencoders for sentiment analysis: a review," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 7, no. 5, p. e1212, 2017.

[29] C. dos Santos and M. Gatti, "Deep convolutional neural networks for sentiment analysis of short texts," in *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, 2014, pp. 69–78.

[30] Y. Kim, "Convolutional neural networks for sentence classification," *arXiv preprint arXiv:1408.5882*, 2014.

[31] P. Ruangkanokmas, T. Achalakul, and K. Akkarajitsakul, "Deep belief networks with feature selection for sentiment classification," in *Intelligent Systems, Modelling and Simulation (ISMS), 2016 7th International Conference on*. IEEE, 2016, pp. 9–14.

[32] Y. Jin, "Deep belief networks for sentiment analysis," Ph.D. dissertation, UNIVERSITY OF NEW BRUNSWICK, 2017.

[33] K. S. Tai, R. Socher, and C. D. Manning, "Improved semantic representations from tree-structured long short-term memory networks," *arXiv preprint arXiv:1503.00075*, 2015.

[34] Y. Suhara, Y. Xu, and A. Pentland, "Deepmood: Forecasting depressed mood based on self-reported histories via recurrent neural networks," in *Proceedings of the 26th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 2017, pp. 715–724.

[35] G. E. Hinton, A. Krizhevsky, and S. D. Wang, "Transforming autoencoders," in *International Conference on Artificial Neural Networks*. Springer, 2011, pp. 44–51.

[36] S. Sabour, N. Frosst, and G. E. Hinton, "Dynamic routing between capsules," in *Advances in Neural Information Processing Systems*, 2017, pp. 3856–3866.

[37] M. Yang, W. Zhao, J. Ye, Z. Lei, Z. Zhao, and S. Zhang, "Investigating capsule networks with dynamic routing for text classification," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 3110–3119.

[38] Y. Wang, A. Sun, J. Han, Y. Liu, and X. Zhu, "Sentiment analysis by capsules," in *Proceedings of the 2018 World Wide Web Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 2018, pp. 1165–1174.

[39] O. Appel, F. Chiclana, J. Carter, and H. Fujita, "A hybrid approach to the sentiment analysis problem at the sentence level," *Knowledge-Based Systems*, vol. 108, pp. 110–124, 2016.

[40] T. Lalji and S. Deshmukh, "Twitter sentiment analysis using hybrid approach," *International Research Journal of Engineering and Technology*, vol. 3, no. 6, pp. 2887–2890, 2016.

[41] G. Montavon, W. Samek, and K.-R. Müller, "Methods for interpreting and understanding deep neural networks," *Digital Signal Processing: A Review Journal*, vol. 73, pp. 1–15, 2018.

[42] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," *arXiv preprint arXiv:1702.08608*, 2017.

[43] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," *ACM Computing Surveys (CSUR)*, vol. 51, no. 5, p. 93, 2018.

[44] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artificial Intelligence*, 2018.

[45] G. Ras, P. Haselager, and M. van Gerven, "Explanation methods in deep learning: Users, values, concerns and challenges," *arXiv preprint arXiv:1803.07517*, 2018.

[46] T. Wang, C. Rudin, F. Velez-Doshi, Y. Liu, E. Klampfl, and P. MacNeille, "Bayesian rule sets for interpretable classification," in *Data Mining (ICDM), 2016 IEEE 16th International Conference on*. IEEE, 2016, pp. 1269–1274.

[47] W. Samek, A. Binder, G. Montavon, S. Lapuschkin, and K.-R. Müller, "Evaluating the visualization of what a deep neural network has learned," *IEEE transactions on neural networks and learning systems*, vol. 28, no. 11, pp. 2660–2673, 2017.

[48] L. A. Hendricks, Z. Akata, M. Rohrbach, J. Donahue, B. Schiele, and T. Darrell, "Generating visual explanations," in *European Conference on Computer Vision*. Springer, 2016, pp. 3–19.

[49] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas *et al.*, "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav)," in *International Conference on Machine Learning*, 2018, pp. 2673–2682.

[50] J. van der Waa, J. van Diggelen, K. v. d. Bosch, and M. Neerincx, "Contrastive explanations for reinforcement learning in terms of expected consequences," *arXiv preprint arXiv:1807.08706*, 2018.

[51] T. Hailesilassie, "Rule extraction algorithm for deep neural networks: A review," *arXiv preprint arXiv:1610.05267*, 2016.

[52] B. Letham, C. Rudin, T. H. McCormick, D. Madigan *et al.*, "Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model," *The Annals of Applied Statistics*, vol. 9, no. 3, pp. 1350–1371, 2015.

[53] H. Lakkaraju, S. H. Bach, and J. Leskovec, "Interpretable decision sets: A joint framework for description and prediction," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. ACM, 2016, pp. 1675–1684.

[54] T. Wang, C. Rudin, F. Doshi-Velez, Y. Liu, E. Klampfl, and P. MacNeille, "Or's of and's for interpretable classification, with application to context-aware recommender systems," *arXiv preprint arXiv:1504.07614*, 2015.

[55] J. Adebayo, J. Gilmer, I. Goodfellow, and B. Kim, "Local explanation methods for deep neural networks lack sensitivity to parameter values," 2018.

[56] M. Sundararajan and A. Taly, "A note about: Local explanation methods for deep neural networks lack sensitivity to parameter values," *arXiv preprint arXiv:1806.04205*, 2018.

[57] A. Shahroudnejad, A. Mohammadi, and K. N. Plataniotis, "Improved explainability of capsule networks: Relevance path by agreement," *arXiv preprint arXiv:1802.10204*, 2018.

[58] D. Doran, S. Schulz, and T. Besold, "What does explainable ai really mean," *A new conceptualization of perspectives. arXiv preprint*, 2017.

[59] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 2625–2634.

[60] R. S. Sutton and A. G. Barton, "Introduction to rl. reinforcement learning: an introduction," 1998.

[61] B. Hayes and J. A. Shah, "Improving robot controller transparency through autonomous policy explanation," in *Proceedings of the 2017 ACM/IEEE international conference on human-robot interaction*. ACM, 2017, pp. 303–312.

[62] A. A. Sherstov and P. Stone, "Improving action selection in mdp's via knowledge transfer," in *AAAI*, vol. 5, 2005, pp. 1024–1029.

[63] J. Clos, N. Wiratunga, and S. Massie, "Towards explainable text classification by jointly learning lexicon and modifier terms," in *IJCAI-17 Workshop on Explainable AI (XAI)*, 2017, p. 19.

[64] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Ng, and C. Potts, "Recursive deep models for semantic compositionality over a sentiment treebank," in *Proceedings of the 2013 conference on empirical methods in natural language processing*, 2013, pp. 1631–1642.

[65] L. Arras, F. Horn, G. Montavon, K.-R. Müller, and W. Samek, "Explaining predictions of non-linear classifiers in nlp," *arXiv preprint arXiv:1606.07298*, 2016.

[66] Y. Zhang, G. Lai, M. Zhang, Y. Zhang, Y. Liu, and S. Ma, "Explicit factor models for explainable recommendation based on phrase-level sentiment analysis," in *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*. ACM, 2014, pp. 83–92.

[67] T. Lei, R. Barzilay, and T. Jaakkola, "Rationalizing neural predictions," *arXiv preprint arXiv:1606.04155*, 2016.

[68] M. Myslín, S.-H. Zhu, W. Chapman, and M. Conway, "Using twitter to examine smoking behavior and perceptions of emerging tobacco products," *Journal of medical Internet research*, vol. 15, no. 8, 2013.

[69] H. Cole-Lewis, A. Varghese, A. Sanders, M. Schwarz, J. Pugatch, and E. Augustson, "Assessing electronic cigarette-related tweets for sentiment and content using supervised machine learning," *Journal of medical Internet research*, vol. 17, no. 8, 2015.

[70] K. Jiang and Y. Zheng, "Mining twitter data for potential drug effects," in *International Conference on Advanced Data Mining and Applications*. Springer, 2013, pp. 434–443.

[71] B. O'Dea, S. Wan, P. J. Batterham, A. L. Calearc, C. Parisb, and H. Christensena, "Detecting suicidality on twitter," *Internet Interventions*, 2015.

[72] D. Mowery, A. Park, M. Conway, and C. Bryan, "Towards automatically classifying depressive symptoms from twitter data for population health," in *Proceedings of the Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, 2016, pp. 182–191.

[73] G. Coppersmith, C. Harman, and M. Dredze, "Measuring post traumatic stress disorder in twitter." in *ICWSM*, 2014.

[74] V. Carchiolo, A. Longheu, and M. Malgeri, "Using twitter data and sentiment analysis to study diseases dynamics," in *Information Technology in Bio-and Medical Informatics*. Springer, 2015, pp. 16–24.

[75] B. O'Dea, S. Wan, P. J. Batterham, A. L. Calear, C. Paris, and H. Christensen, "Detecting suicidality on twitter," *Internet Interventions*, vol. 2, no. 2, pp. 183–188, 2015.

[76] J. C. Eichstaedt, H. A. Schwartz, M. L. Kern, G. Park, D. R. Labarthe, R. M. Merchant, S. Jha, M. Agrawal, L. A. Dziurzynski, M. Sap *et al.*, "Psychological language on twitter predicts county-level heart disease mortality," *Psychological science*, vol. 26, no. 2, pp. 159–169, 2015.

[77] H.-J. Kim, S.-B. Park, and G.-S. Jo, "Affective social network—happiness inducing social media platform," *Multimedia Tools and Applications*, vol. 68, no. 2, pp. 355–374, 2014.

[78] A. Holzinger, C. Biemann, C. S. Pattichis, and D. B. Kell, "What do we need to build explainable ai systems for the medical domain?" *arXiv preprint arXiv:1712.09923*, 2017.

[79] A. Panchenko, E. Ruppert, S. Faralli, S. P. Ponzetto, and C. Biemann, "Unsupervised does not mean uninterpretable: The case for word sense induction and disambiguation," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, vol. 1, 2017, pp. 86–98.