ContextMEL: Classifying Contextual Modifiers in Clinical Text

ContextMEL: un Clasificador de Modificadores Contextuales en Texto Clínico

Paula Chocrón, Álvaro Abella, Gabriel de Maeztu
IOMED
{paula.chocron,alvaro.abella,gabriel.maeztu}@iomed.es

Abstract: Taking advantage of electronic health records in clinical research requires the development of natural language processing tools to extract data from unstructured text in different languages. A key task is the detection of contextual modifiers, such as understanding whether a concept is negated or if it belongs to the past. We present ContextMEL, a method to build classifiers for contextual modifiers that is independent of the specific task and the language, allowing for a fast model development cycle. ContextMEL uses annotation by experts to build a curated dataset, and state-of-the-art deep learning architectures to train models with it. We discuss the application of ContextMEL for three modifiers, namely Negation, Temporality and Certainty, on Spanish and Catalan medical text. The metrics we obtain show our models are suitable for industrial use, outperforming commonly used rule-based approaches such as the NegEx algorithm.
Keywords: Clinical text, Temporality, Negation, Certainty, deep learning, annotation

Resumen: Las historias clínicas electrónicas pueden traer grandes avances en la investigación médica, pero requieren el desarrollo de herramientas para procesar texto no estructurado en diferentes idiomas. Una tarea clave es la detección de distintos modificadores contextuales, como el aspecto temporal de un concepto, o si está negado. En este trabajo presentamos ContextMEL, un método para construir clasificadores para modificadores contextuales que es independiente tanto de la tarea específica como del lenguaje, permitiendo un ciclo de desarrollo dinámico. ContextMEL usa anotaciones de expertos para crear un dataset curado, y las últimas tecnologías en aprendizaje profundo. En este artículo discutimos la aplicación de ContextMEL para tres modificadores contextuales (temporalidad, negación, y certeza) en texto médico en castellano y catalán. Los resultados obtenidos muestran que nuestros modelos pueden utilizarse en un entorno industrial, y que son más precisos que conocidos métodos basados en reglas, como el algoritmo NegEx.
Palabras clave: Texto clínico, clasificación, aprendizaje profundo, anotación

1 Introduction

The growing adoption of electronic health records (EHR) in hospitals provides a unique opportunity to analyze clinical text data automatically. Firstly, being able to efficiently and accurately extract data from these unstructured text corpora would greatly accelerate the analysis process. Secondly, it would make it easier to comply with the privacy policies that such sensitive text requires. Driven by these considerations, Natural Language Processing (NLP) applications for the clinical domain has been a very active field of research in the past decades.

The main goal of our NLP system is to find the units of observation (i.e data points) required by clinical research from the EHR of the patients. Therefore extracting data points from clinical text usually means, at its core, understanding whether a given written observation is true or false for each case. For example, a study may require the identification of all the cases of anemia, to later analyze pathologies that are commonly shared with that diagnosis. This requires as a first step a tool to identify concepts in a text, commonly known as a *concept extractor*. These tools standardly take advantage of some of the large and curated knowledge bases that exist for medical terminology such as SNOMED CT[1], performing matches between their entities and the ones that are found in the text. These matches can either use syntactic distances or rules (Abacha and Zweigenbaum, 2011), or machine learning (Torii, Wagholikar, and Liu, 2011).

To construct an appropriate dataset, however, it is also essential to be able to identify contextual modifiers that modify the concepts in question. For example, consider the following text: *Previous episodes of Anemia, the patient's Hgb levels are now normal.* Although a good concept extractor would find a reference to *anemia*, it is as important to determine that it refers to the past, and that it is therefore not an active pathology anymore. Another typical case in which contextual modification is pivotal is in the identification of whether a concept is negated or not. In the next section we discuss different approaches which have been proposed to address these kind of problems. These tools are, in general, language-specific, and designed to identify one specific contextual modifier (such as Negation). However, the field of clinical research is constantly changing, and it is very likely to have studies that will need to handle modifiers that have never been taken into account before.

In this paper we present ContextMEL, a system that allows to train models to perform contextual concept analysis, irrespective of their source language and the particular task. At its core, ContextMEL proceeds by annotating data and then training models with these annotations. Here, we propose to use established state-of-the-art deep learning models that are particularly well-suited to solve the problem of assigning labels to a concept in the context of a longer text. These models are trained and evaluated on three different context-related tasks, all applied to clinical data, using both Spanish and Catalan sources. Since our method and the deep learning models we use are domain and language independent, our method can be straightforwardly applied to other domains or languages. Our technique shows promising results for these tasks, and it outperforms one of the best-known methods for Negation detection.

[1]http://www.snomed.org/

2 Related Work

The problem of identifying contextual concept modifiers is very broad, and it has been researched extensively. We will restrict to discuss approaches developed for medical text. Most of the existent approaches tackle the detection of a particular contextual modifier, and for a specific language. We will discuss solutions that identify Negation, Certainty and Temporality.

The development of tools to determine contextual concept modifiers for medical text has evolved side by side with general progress in NLP techniques. Early approaches relied on syntactic and rule-based methods. For example, the NegEx algorithm (Chapman et al., 2001) applies a rule-based system to perform Negation identification in English. NegEx has been extended to include other modifiers in systems such as ConText (Harkema et al., 2009), which can also identify the experiencer (i.e. the patient or a member of her family) and a concept's temporal modifier. NegEx has also been extended to other languages. Relevant examples for Spanish are (Costumero et al., 2014; Stricker, Iacobacci, and Cotik, 2015) and `NegEx-MES`, which we will discuss later.

More recent techniques incorporate the use of machine learning methods to the problem of contextual modifier identification, such as Conditional Random Fields (Agarwal and Yu, 2010). The work in (Cruz Diaz et al., 2012) uses machine learning techniques to identify Negation and Speculation (Certainty) for Spanish medical text. Finally, the latest approaches rely on deep learning techniques. The work in (Dalloux, Claveau, and Grabar, 2019) uses Long-Short Term Memory (LSTM) networks to solve Negation and Certainty in French corpora, while (Fancellu, Lopez, and Webber, 2016) applies them for English. The work on (Tourille et al., 2017) uses them to detect temporal modifiers in English. Finally, the latest work on modifier detection for clinical text uses modern language models based on Transformers such as BERT. Examples are (Khandelwal and Sawant, 2019) for Negation in English, (García-Pablos, Perez, and Cuadros, 2020) for Negation in Spanish, and (Sergeeva et al., 2019), which also tackles Speculation.

3 *Context-MEL: a General System for Contextual Classification*

ContextMEL is a system to tackle the problem of *contextual modifier analysis*. This problem can be defined formally as follows: Let t be a text, and c be a concept formed by a subset of one or more words in t. Given a set of labels L, assign a label from L to c in the context of t. For example, in the problem of identifying negations, the set of labels is composed of $L = \{\texttt{Positive, Negative}\}$. Given $t =$ *"The patient does not have anemia."* and the concept $c =$ *anemia*, the objective of the task is to assign one of the two labels to c in t. In theory, concepts could be any n-gram. In practice, we will use (and therefore develop methods for) only those that are recognized as medical concepts by a knowledge base.

ContextMEL provides a general framework to build classification models that can be used to solve the problem of contextual modifier analysis for different tasks, independently of the language of the text or its domain. Concretely, the system consists of a first phase, during which it uses manual expert annotators to build a labeled dataset for training. In a second phase we use this dataset to train a classification model that can assign labels to concept-text pairs. We use well-known model architectures that were originally developed for the problem of Aspect-Based Sentiment Analysis, and we show that we can apply them without changes to a variety of other tasks.

The models that are obtained with ContextMEL can be applied directly on a concept extraction pipeline. First, texts are preprocessed and analyzed by a concept extractor which links expressions to entities in a knowledge base. Then, these concepts together with the original text are fed to the contextual models built with ContextMEL, which assigns them a label. The methods that conform our concept extractor are outside of the scope of the paper, and therefore are not discussed in depth here.

ContextMEL assumes the following resources are available: 1. A large corpora of unlabeled text. 2. A relatively basic system that finds *concepts*, or the entities that are of interest and would require contextual labels. 3. A team of expert annotators.

The first and the second items are normally available when working with clinical data. Unlabeled text abounds, and there are reliable, well-curated knowledge bases that can be useful to identify concepts. In our case, we used a simple NER system built on ULMS vocabulary, with most of the terms coming from the Spanish SNOMED CT (see (Torii, Wagholikar, and Liu, 2011) and (Soldaini and Goharian, 2016) for examples of other medical concept extraction systems). Annotators, on the other hand, are normally expensive. However, we will show that a relatively small amount of annotated data is necessary.

Our system aims to be as generalizable as possible, providing a procedure that can be adapted to different domains and contextual modifiers that need to be identified. In our particular case, we used the system to build models to classify concepts in medical text written in Spanish and Catalan. In many cases one same text contains bits of both languages. We used the framework to build models for three different contextual modifiers:

Temporality. The goal of the Temporality task is to understand whether a concept belongs to the *current episode*, to the *clinical history* or to the *future plan* of a patient. The set of labels for this modifier is $\{$`Antecedent, Plan, Current Episode`$\}$. For example, in *"Patient with kidney transplant in 2010. Has mild back pain. I prescribe Ibuprofen every 8 hours."*, the concept *kidney transplant* is `Antecedent`, *back pain* is `Current Episode`, and *Ibuprofen* is `Plan`.

Certainty. This task is about identifying whether a concept is certain, or if it refers to a hypothesis or conditional that could be true or not. The set of labels for this modifier is $\{$`Certain, Uncertain`$\}$. For example, in *"If there is high fever, take ibuprofen"*, the concept *fever* is a hypothesis, while *ibuprofen* is conditional (depends on the fever). They both are, therefore, uncertain. It could be argued that no concept is totally certain, since the doctor could be wrong. It is important to keep in mind that we are not trying to classify whether a concept is true or not, but rather if the doctor considers it as a fact. Certainty is sometimes referred to as *speculation* in the literature.

Negation. Negation is perhaps the simplest of the tasks. It aims to identify whether a concept is negated or not in a text. The set of labels for this modifier is {`Positive, Negative`}. For example, in the sentence *"Patient with anemia had no liver abnormality."*, *anemia* is positive and *liver abnormality* is negative. We have already discussed related work for this task in the introduction.

4 *Annotation: Building a training dataset for contextual modifier analysis*

The first step of Context-MEL consists on producing a high-quality dataset that can be used for training a machine learning model. While the resulting datasets are specific to a particular contextual modifier, language, and domain, we provide a general framework to obtain annotations.

Suppose we are trying to solve a *contextual modifier analysis* problem with set of labels L. The annotation phase requires to employ a team of domain expert annotators, who will assign labels from L to pairs of text-concept. To do so, annotators have access to an interface where they can read a text with a highlighted concept. The highlighted concepts are obtained automatically from the text, using the automatic *concept matching* system we mentioned in the previous section. Next to the text there is a table where they can select labels from L (whether it is possible to select more than one label depends on the specific semantics of the contextual modifier being analyzed). Finally, annotators can choose to finish the annotation of the concept by pressing "OK" (meaning that they submit their labels), "NO" (meaning that there is something wrong with the concept) or "PASS" (meaning that they are unsure of the annotation and they prefer to jump to another concept). Once a concept is annotated, another one from the same text is highlighted, until the annotator labels them all and jumps to the next text.

In our particular case, we employed a team of five medical practitioners. To make annotation more efficient, and since we were developing models for the three modifiers (*Temporality, Certainty, Negation*) in parallel, we merged their labels to create one unique annotation task. We only included labels that were not the default one for each of the modifiers: `History, Plan, Uncertain` and `Negative`. Since we were mixing independent modifiers, we made labels not exclusive, and annotators could choose more than one. For example, they could mark the word *ibuprofen* in *"take ibuprofen if there is pain"* as `Plan, Uncertain`. Figure 1 shows an example of a concept annotation.

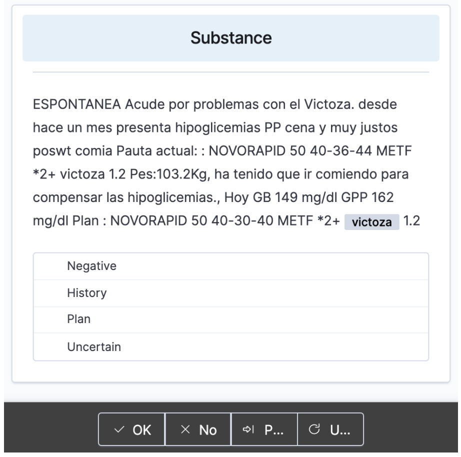

Figure 1: Labels for the annotator's task

The full annotation cycle lasted 3 months. We organized a meeting in person at the beginning and one midterm, to explain the task and clarify possible doubts.

We gave each pair concept-text to three different annotators, varying the composition of the groups. We considered an annotation as valid if at least two of the three annotators agreed on the label.

4.1 A more efficient annotation strategy: Confirming Entities First

We noticed rapidly of a problem in our approach: it relied too much on the automatic concept matcher, which was not perfect. In particular, our matcher had many false positives that were considered concepts although they should not be. This meant that annotators had to spend valuable time deciding how to label something that was not really a concept worth labeling. To tackle this problem, we set up a preliminary task that consisted on confirming whether a sequence of words was a medical concept or not. Annotators where very fast on this task, since they only had to choose between two labels. Later, we used these results on the main task, by assigning them only those concepts that had been verified as correct.

5 *Training: building classifiers for contextual modifier analysis*

Once we obtained the annotated dataset, we proceeded to train a model that would perform the labeling automatically. We propose to consider the problem of contextual modifier analysis as a generalization of *aspect-based sentiment analysis (ABSA)*. ABSA is a type of sentiment analysis that takes into account the fact that a text can express different sentiments about different things. For example, the text *"The movie has a very interesting plot, but I found the main actor bland"* is not positive or negative per-se, but rather positive about the plot and negative about the actor. ABSA can be defined as classifying the polarity of a concept c (which they call *aspect*) in a context t. This is equivalent to our problem when the set of labels L express polarity (for example `Positive, Negative`). The community working on solving ABSA has proposed different architectures to perform this kind of targeted classification; a survey can be found in (Schouten and Frasincar, 2015). For ContextMEL we compared the performance of an approach based on LSTMs and one based on BERT.

	Temp	Certainty	Negation
Train	6286/1000	1743/282	4931/689
Val	1459/255	384/84	1162/ 218
Test	261/79	246/90	173/75

Table 1: Size of the datasets, with the number of Spanish/Catalan examples on each one

Before discussing the models in the following subsections, let us comment on the datasets we used for training. Our data was naturally highly imbalanced (for example the *Temporality* task was skewed to the "Current Episode" label). To minimize the effect of this bias, we used balanced datasets that contained 65% of the prevalent class and 35% of the other one for the datasets with two labels (*Certainty* and *Negation*) and a $40\% - 20\% - 20\%$ split for the *Temporality* dataset. Table 1 shows the dimensions of the train and validation datasets that we used for training, together with a testset that we will discuss in the next section. As we can see, the datasets are mixed in Spanish and Catalan, although Catalan is notably underrepresented. Note that the datasets are not very large. For Certainty, for example, training with only 2025 examples (of which just 709 are uncertain) is enough to achieve the results we report next.

5.1 Contextual analysis via LSTM models

The first alternative we explored was to use a method proposed in (Tang et al., 2016). Their approach consists essentially in solving ABSA with two unidirectional LSTM networks, one modeling the part of the context to the left of the concept, and the other one the part to the right. The LSTM that models the right part receives the text reversed, so that the input to both networks ends with the concept. More formally, if the context c has n tokens and the concept to analyze is in positions i to $i + t$, the left LSTM receives input $c[0 : i+t]$ and the right LSTM receives $reverse(c[i : n])$. Using a concrete example, if the we want to classify the concept *anemia* in the text *"60 years old woman. History of anemia. Normal values now."*, the inputs for the left and right LSTMs would be:

Left: `60 years old woman . History of anemia`
Right: `. now values Normal . anemia`

These models are sometimes referred to as TD-LSTM (*Target-Dependent LSTM*).

We trained an LSTM with 1 layer for each of our context dimensions. To vectorize the input, we used a specialized embedding that we trained on our large corpora of unlabeled Spanish and Catalan texts. This embedding had 200 dimensions and was trained using word2vec (Mikolov et al., 2013). We trained the LSTMs for 15 epochs on each dataset. We were able to train these models on a machine with only one CPU and 8 GB of RAM memory.

5.2 Contextual analysis via BERT

The second approach we investigated was a BERT-based methodology. Language models such as BERT (Devlin et al., 2018) are neural networks based on the Transformer architecture (Vaswani et al., 2017) which have been trained on a huge dataset for a general task such as guessing a masked word. This setup makes them very flexible, which has proved very beneficial to perform transfer learning. In other words, the knowledge in the language model can be reused to solve different specific tasks, needing only a small task-specific dataset.

BERT architectures have been used in different ways to solve ABSA (Zeng et al., 2019). In our case we used a very simple approach that took advantage of the type of input expected by BERT. In addition to the masked word task that we already mentioned, BERT is also trained for the *next sentence prediction* task. Shortly, the task consists on deciding whether two sentences are consecutive or not. To implement this task, the model uses a special token `[SEP]`, which indicates the separation between two sentences. If the input sentences are a and b, the model input is then [`[CLS]` a `[SEP]` b], where `[CLS]` is a special token used at the beginning of text. We used this mechanism to give, as input to the model, both the text and the concept, using `[SEP]` to indicate their boundaries. That is, for a pair of text t and concept c, our input was [`[CLS]` t `[SEP]` c]. Using the same example as before, to classify *anemia* in *"60 years old woman. History of anemia. Normal values now."* the input would be [`[CLS] 60 years old woman . History of anemia . Normal values now . [SEP] anemia [SEP]`].

Since our text is in Spanish and Catalan, we used the Multilingual (*bert-base-multilingual-cased*) version of BERT as a base, which we fine-tuned with the same datasets we used for the LSTM-version[2]. The models required only 3 epochs to achieve their best performance. To train the BERT models, we used a machine with 1 GPU and 30 GB RAM.

6 Results

In this section we present and discuss the results we obtained using both techniques (LSTM and BERT-based) for the three tasks we described before. To evaluate our models as accurately as possible, we built a golden testset for each task which was manually revised by us, using the annotator's data as a basis. For each label l (including a label `Blank` as a common one for `Positive, Current Episode, Certain`), we selected 100 examples that at least 2 annotators had marked with l, and 100 examples where only one one annotator had marked it as such (which often consist on harder or edge cases). Then we labeled the examples ourselves, and used the results as a golden test dataset.

To ensure a fair evaluation framework, we used in the testset only notes that had not been seen at train time. That is, we not only made sure the pair concept-text was previously unseen, but we also used completely new texts. Table 2 shows the results for the three different tasks. As we can see, we obtain good accuracy levels, which could be used industrially. Results are particularly good for the Negation task, which is unsurprising since it is notably simpler than the other two. Using BERT models improves the accuracy significantly for Tense and Certainty, which is expected since it is a very large model. When evaluating the results, it is important to keep in mind that training BERT requires a GPU and more memory, while the LSTMs can be trained on a much simpler computer; the size of the model would also impact the inference time. Notably, BERT models have better results even when they use the very generic Multilingual model as a base, while the LSTM ones use a vector space specifically designed for the task. Investigating whether further training the base model leads to better results is part of our planned future research.

In Table 3 we also show the precision ans recall values for each class. As we can see, the *Temporality* case is particularly good identifying future plans, which are normally very clear in the text, while the difference between Current Episode and Antecedent is sometimes more subtle. The Negation model is, as expected, better at identifying positives (the prevalent class). In Certainty, we observe the certain case is slightly better.

	Temp	Certainty	Negation
LSTM	81.82%	85.71%	95.70%
BERT	84.41%	88.10%	95.16%

Table 2: Accuracies for the three tasks for each model

We consider the possibility of building models specifically designed for a particular language, task, and domain to be one of the main advantages of our system. However, this makes comparing our techniques

[2]While there are Spanish-only versions of BERT, we are not aware of any Catalan-Spanish one, or even any Catalan-only version. Catalan is, however, included in the multilingual version of BERT

	Precision	Recall
Certain	85.06%	87.06%
Uncertain	86.42%	84.34%
Positive	96.67%	95.47%
Negative	92.86%	94.71%
Past	77.93%	78.03%
Present	77.89%	77.93%
Future	89%	88.80%

Table 3: Precision and recall per class for each task (LSTM)

NegEx	Context-MEL
83.14%	94.22%

Table 4: Results for the Spanish dataset

against other tools challenging. To the best of our knowledge, there are no available tools to perform temporal analysis of clinical text in Spanish and Catalan. The task of identifying negations has been the most prolific of the three we worked on. In what follows we show how our model compares with the `NegEx-MES`[3] system, which applies the `NegEx` technique to Spanish medical texts. `NegEx` is, at this moment, the best-known method to identify negated concepts, and its implementations are commonly applied industrially.

Since NegEx-MES was developed for Spanish texts, we compared the systems only for that portion of our testset (a 73.59%, according to Table 1). These results can be seen in table 4. As we can see, our system improves significantly the accuracy over NegEx-MES.

7 Conclusions and Future Work

We described ContextMEL, a system to build, in a general way, models to detect contextual modifiers. The framework includes an efficient way of annotating data by experts and state-of-the-art model architectures to harness the knowledge in the resulting dataset. Importantly, the development of the system is not bounded to a particular language or a particular task, allowing for a fast, simple, and organized model development cycle.

We applied ContextMEL for the identification of three context modifiers: Temporality, Negation, and Certainty. We obtained promising results that are already, after only one development iteration, suitable for industrial use. Moreover, the results for Negation outperformed the ones obtained with a commonly used rule-based tool. In summary, we think ContextMEL harnesses the power of the latest deep learning architectures to make the best possible use of a dataset that is carefully built, but not huge. For example, for Certainty we obtained an accuracy of 88.1% with only 2025 (and 709 uncertain) examples.

We have two clear directions of development and research for this project. First, we plan to study whether using a BERT model specially trained for Spanish and Catalan clinical text as a base for our fine-tuning would improve the results we already have. This model does not exist yet, so we plan to build it ourselves with our corpora of unlabeled data. If results are better, it would mean another step towards building a general framework to make fine-tuning for specific tasks even more lightweight.

As another line for future work we plan to work on a Quality Assessment-based development cycle to refine models using annotations. Our objective is to define a workflow that allows us to build the training dataset iteratively and on demand. This will be achieved by evaluating systematically the models that are obtained with each iteration until they reach an industry-suitable accuracy threshold. This avoids annotating unnecessarily large amounts of data, which is yet another way of making model development faster.

References

Abacha, A. B. and P. Zweigenbaum. 2011. Automatic extraction of semantic relations between medical entities: a rule based approach. *Journal of biomedical semantics*, 2(S5):S4.

Agarwal, S. and H. Yu. 2010. Biomedical negation scope detection with conditional random fields. *Journal of the American medical informatics association*, 17(6):696–701.

Chapman, W., W. Bridewell, P. Hanbury, G. Cooper, and B. Buchanan. 2001. A simple algorithm for identifying negated

[3]https://github.com/PlanTL-SANIDAD/NegEx-MES

findings and diseases in discharge summaries. *Journal of Biomedical Informatics*, 34:301–310, 11.

Costumero, R., F. Lopez, C. Gonzalo-Martín, M. Millan, and E. Menasalvas. 2014. An approach to detect negation on medical documents in spanish. In *Brain Informatics and Health*. Springer International Publishing, pages 366–375.

Cruz Diaz, N., M. López, J. Mata Vázquez, and V. Pachón. 2012. A machine-learning approach to negation and speculation detection in clinical texts. *Journal of the American Society for Information Science and Technology*, 63:1398–1410, 07.

Dalloux, C., V. Claveau, and N. Grabar. 2019. Speculation and negation detection in French biomedical corpora. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 223–232, Varna, Bulgaria, September. INCOMA Ltd.

Devlin, J., M. Chang, K. Lee, and K. Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Fancellu, F., A. Lopez, and B. Webber. 2016. Neural networks for negation scope detection. In *Proceedings of the 54th annual meeting of the Association for Computational Linguistics (volume 1: long papers)*, pages 495–504.

García-Pablos, A., N. Perez, and M. Cuadros. 2020. Sensitive data detection and classification in spanish clinical text: Experiments with bert. *arXiv preprint arXiv:2003.03106*.

Harkema, H., J. N. Dowling, T. Thornblade, and W. W. Chapman. 2009. Context: An algorithm for determining negation, experiencer, and temporal status from clinical reports. *Journal of biomedical informatics*, 42 5:839–51.

Khandelwal, A. and S. Sawant. 2019. Negbert: A transfer learning approach for negation detection and scope resolution.

Mikolov, T., K. Chen, G. Corrado, and J. Dean. 2013. Efficient estimation of word representations in vector space.

Schouten, K. and F. Frasincar. 2015. Survey on aspect-level sentiment analysis. *IEEE Transactions on Knowledge and Data Engineering*, 28:1–1, 01.

Sergeeva, E., H. Zhu, A. Tahmasebi, and P. Szolovits. 2019. Neural token representations and negation and speculation scope detection in biomedical and general domain text. In *Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019)*, pages 178–187.

Soldaini, L. and N. Goharian. 2016. Quickumls: a fast, unsupervised approach for medical concept extraction. In *MedIR workshop, sigir*, pages 1–4.

Stricker, V., I. Iacobacci, and V. Cotik. 2015. Negated findings detection in radiology reports in spanish: an adaptation of negex to spanish. 07.

Tang, D., B. Qin, X. Feng, and T. Liu. 2016. Effective LSTMs for target-dependent sentiment classification. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3298–3307, Osaka, Japan, December. The COLING 2016 Organizing Committee.

Torii, M., K. Wagholikar, and H. Liu. 2011. Using machine learning for concept extraction on clinical documents from multiple data sources. *Journal of the American Medical Informatics Association*, 18(5):580–587.

Tourille, J., O. Ferret, A. Neveol, and X. Tannier. 2017. Neural architecture for temporal relation extraction: a bi-lstm approach for detecting narrative containers. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 224–230.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. 2017. Attention is all you need.

Zeng, B., H. Yang, R. Xu, W. Zhou, and X. Han. 2019. Lcf: A local context focus mechanism for aspect-based sentiment classification. *Applied Sciences*, 9:3389, 08.