Leveraging knowledge sources for detecting self-reports of particular health issues on social media

Parsa Bagherzadeh and Sabine Bergler
CLaC Labs, Concordia University
Montreal, Canada
{p_bagher, bergler}@cse.concordia.ca

Abstract

This paper investigates incorporating quality knowledge sources developed by experts for the medical domain as well as syntactic information for classification of tweets into four different health oriented categories. We claim that resources such as the MeSH hierarchy and currently available parse information are effective extensions of moderately sized training datasets for various fine-grained tweet classification tasks of self-reported health issues.

1 Introduction

Social media are a ubiquously accessible way to communicate and interact with others, making their users producers of Big Data at a fast rate. It is estimated that about 500M tweets are sent each day on Twitter which often contain information about opinions, trends, reviews, health, incidents, etc. This offers the possibility to gain insight into individuals' behavior and general state in direct and unmitigated fashion (Rousidis et al., 2020).

Health applications based on social media are an active research area for outbreak management, disease surveillance (Charles-Smith et al., 2015), and pharmacovigilance (Golder et al., 2015). For instance, epidemiologists hope to mine social media to predict and monitor the likelihood and possible severity of outbreaks in a timely fashion. Systems that support this type of research have to make predictions from incomplete data of varying quality.

Deep learning methods are popular for NLP applications and demonstrated significant improvements in areas such as text classification. Deep models have been widely used for personal health mention detection (Khan et al., 2020), (Sarabadani, 2019), (Barry and Uzuner, 2019), (Aroyehun and Gelbukh, 2019), (Bagherzadeh et al., 2018). Deep models, however, do not usually have access to outside resources, apart from word embeddings. While such models can outperform systems that are limited to look-up in gazetteer lists for task specific terms, this can only be when the terms of the test set are foreshadowed sufficiently in the training set.

Sensitivity to lexical triggers is crucial in classification, especially in the medical domain, where vocabularies are ever-growing and new specialized terms are introduced everyday. The most recent example is the term "CoVID" which was coined in late 2019.

Most language models are trained contextually. The assumption for contextualized language models is that the meaning of a word can be represented by the context in which it appears. However, the context usually is not sufficient to represent the meaning for rare specialized terms, which require large amounts of training data for coverage. In addition, highly specialized terms with very different meanings may occur in the same immediate context (see Example 1), rendering contextualized word embeddings less effective.

(1) (a) *My son was diagnosed with leukemia*
(b) *My son was diagnosed with hydrocephalus*

The context for *hydrocephalus* and *leukemia* here is the same, and is the same for all diseases, making such contextualized word embeddings less sensitive to, for instance, the more fine-grained distinctions between *birth defects* and *cancer*. Consequently, contextualized language models often fail to make these distinctions.

Current language models have over 60M parameters[1], making fine-tuning as well as testing time-consuming and requiring large training sets.

[1] The smallest model, DistilBERT (Sanh et al., 2019), has 66M parameters, RoBERTa$_{Large}$ (Liu et al., 2019) has 340M parameters.

Proceedings of the 12th International Workshop on Health Text Mining and Information Analysis, pages 38–48
April 19, 2021. ©2021 Association for Computational Linguistics

These issues motivate us to investigate widely available external knowledge sources, such as MeSH (Lipscomb, 2000), and language features in a deep architecture suitable for personal health mention detection. We show that knowledge sources, combined with light-weight word embeddings and language models such as GLoVE (Pennington et al., 2014) and ELMo (Peters et al., 2018), are strong contenders for larger models such as RoBERTa (Liu et al., 2019).

We experiment on four health-related tweet classification tasks of the ongoing SMM4H Workshop and present ablation studies to assess the contribution of different external knowledge sources. Our results suggest that the external resources tested indeed enhance performance when properly calibrated to work together. Best performance is achieved with a two layer system that adds representations of gazetteer lists and enhanced part-of-speech annotations in an encoder followed by a graph convolutional neural network (GCNN) (Kipf and Welling, 2017) representing preprocessed grammatical dependencies.

2 Related literature

Personal experiences posted on social media can give insight into the state of public health. Examination and identification of smoking behavior (Myslín et al., 2013), non-medical use of opioids (Chan et al., 2015), and identification of medication-related experiences (Jiang et al., 2018), (Jiang et al., 2019) have recently been studied on social media.

A variety of models have been proposed for the task addressed in the current paper, namely health experience mention detection for different experiences. The approaches fall into three main categories, namely statistical models with hand-crafted features, pure deep learning models, and deep models with leveraged features.

Hand-crafted features (Jiang et al., 2016) proposed a set of textual features such as count of emotion words, of unique words, of first person pronouns, of pronouns, etc., for personal health surveillance. (Jiang et al., 2019) compared different word embeddings such as GLoVE, Word2Vec (Mikolov et al., 2013), fastText, and wordRank with the features of (Jiang et al., 2016). Their word embeddings performed close to one another and considerably outperform their feature based model.

For detection of vaccination behaviour, (Joshi et al., 2018) proposed to use the count of POS tags, number of special characters (such as # and @), and count of emotion words as input features to an ensemble of SVM, logistic regression and random forest classifiers. (Joshi et al., 2018) also experimented with a pure deep-learning method (employing ULMfit (Howard and Ruder, 2018) with fine-tuning) and reported a performance close to their feature-based model, demonstrating that a model with handcrafted features is a strong contender for deep models.

To identify drug and adverse drug reaction mentions, (Saha et al., 2018) used a SVM classifier with some hand-crafted features such as the count of typed-dependency relation, drug names, and sentiment score and demonstrated success to some extent. (Çöltekin and Rama, 2018) also addressed the task using a SVM model with word and character n-grams as input features. Bag of word features (tf-idf) as well as negation, adverse reaction mentions, and drug mention were also used in (Wang et al., 2019) as input for a SVM.

Models with hand-crafted features have shown competitive performance compared to deep models on some tasks and datasets but despite the availability of many high quality resources from which features can be derived, the power of contextualized word embeddings to fill in not only lexical gaps, but also subtask specific patterns led to the investigation of deep models.

Pure deep models A majority of the proposed models for personal health detection are deep learning based. Studies such as (Xherija, 2018) and (Cortes-Tejada et al., 2019) use conventional pre-trained word embeddings, such as Word2Vec (Mikolov et al., 2013) and GLoVe. By the advent of pre-trained language models, many studies benefit from models such as ULMfit (Howard and Ruder, 2018), ELMo (Peters et al., 2018), and BERT (Devlin et al., 2019) variants (Khan et al., 2020), (Sarabadani, 2019), (Miftahutdinov et al., 2019), (Aroyehun and Gelbukh, 2019), (Babu and Eswari, 2020), (Aduragba et al., 2020).

Deep models are dependent on the representativeness of their training data for the test cases. Pretrained (often BERT-based models) are generally the top performers in current shared task competitions. The models have to be developed by highly skilled machine learning experts. The data, on the other hand, have to be collected and annotated by domain experts, if they are to be of use in health care research.

Deep models with leveraged features While many approaches to health-oriented classification of tweets use features in statistical models, there have been fewer efforts to leverage them in deep models.

For the task of adverse drug reaction mention detection (Wu et al., 2019) proposed to embed POS tags, gazetteer information, and sentiment scores, then concatenate the features to GloVe embeddings as input to a hybrid of CNN and LSTM. POS tags as well as features such as side effects, medical concepts, and first character are concatenated to word embeddings in (Vydiswaran et al., 2019). (Bagherzadeh et al., 2019) also leveraged features such as adverse mentions, POS tags, and scope of negation and modality, by concatenation to Word2Vec and GLoVE embeddings.

The limitation of input feature concatenation is that only a fixed number of annotation types can be used. Adding more gazetteer lists requires reconfiguring the network, since the number of dimensions is bounded by the number of predefined annotations. This is undesirable because addition of any annotation type has to be performed by a machine learning expert and must be followed by re-training the network from scratch.

In the following we outline a simple architecture, where the end-user (possibly an epidemiologist) can add or remove gazetteer lists and fine-tune the model without making any changes in the network settings such as hidden dimensions (and thus without requiring help from a machine learning expert).

3 Tasks

In order to demonstrate the ability of the presented approach to adapt to new domains, we compare performance on four tasks from the Social Media Mining for Health application[2] (SMM4H) shared tasks. All tasks involve detection of self-reported health mentions on Twitter.

SM18-2: Self-reported medication intake is a 3-class. Tweets which clearly express personal medication intake are considered Category 1. Tweets where the user *may* have taken some medication are Category 2. Category 3 tweets mention medication names but do not indicate personal intake (Weissenbacher et al., 2018).

(2) (Class 1):
*I took three Ibuprofens and I still got a headache crack head *cough cough**

(3) (Class 2):
since I'm constantly in pain, the only way I can go to sleep is if I take Tylenol PM

(4) (Class 3):
Will you take a Xanax and relax.

The performance is evaluated as μF score of Class 1 and Class 2.[3]

SM18-4 Vaccination behavior mention classification is a binary task where the positive class indicates the user has received or intends to receive a flu vaccine. A tweet is classified as negative if it does not contain any mentions of a vaccination or if it merely mentions vaccination (Weissenbacher et al., 2018).

(5) (Class 0):
scientists found a flu vaccine flaw, now they have to fix it

(6) (Class 1):
waiting at the pharmacy for my flu shot

SM19-1 Adverse drug reaction mention is the task of identifying mentions of side effects as the results of drug consumption (Weissenbacher et al., 2019).

(7) (Class 0):
I'm so proud of bob for taking xarelto!

(8) (Class 1):
This Vyvanse got me sweating right now and I dont even know why

SM20-5 Birth defect mention detection is a 3-way classification problem, where Category 1 tweets refer to the user's child and indicate that he/she has a birth defect. Category 2 tweets are unclear whether the tweet speaks of birth defects of the author's child. Category 3 tweets merely mention birth defects but not with respect to the author's child (Klein et al., 2020). Examples of each class are provided in Examples 9-11.

(9) (Class 1):
I had a stillbirth when I was 7 month pregnant. It was hydrocephalus.

[2] https://healthlanguageprocessing.org/smm4h-2021/

[3] for all tasks we follow the standard measure used in SMM4H competitions

(10) (Class 2):
Olivia was born with down syndrome.

(11) (Class 3):
Down's syndrome day. Please share to raise awareness.

The performance is evaluated as the μF score of Class 1 and Class 2.

A summary of the statistics of training and test data is provided in Table 1.

Table 1: Statistics of the data sets

Task	Train	Test
SM18-2	14219	3554
SM18-4	4579	1144
SM19-1	25678	4575
SM20-5	18382	4603

4 External resources

We experiment with two types of external knowledge sources for the deep learning system: (a) gazetteer lists extracted from MeSH as examples of a high quality resource developed by experts that can be used to partly define the domain of the task and (b) language features (POS, NEs, dependencies) extracted from the text with a parser and named entity recognition pipeline.

4.1 Gazetteer lists

Disease Mentions of disease are important evidence for medication intake classification, since drugs are usually consumed to treat a disease or its symptom. To identify disease mentions, we compiled a gazetteer from subtree C in MeSH (Lipscomb, 2000) which includes terms for *infections, wounds, injuries, pain, etc.*

(12) *: I've literally had a headache all day today and have taken four Tylenols throughout the day !*

Drug To identify drug mentions, we use the DrugBank database (Wishart et al., 2018), which includes commercial drug names as well as their scientific names.

Anatomy Body parts are often present in both birth defect and adverse drug reaction mentions. Tweets talk about a child's birth defect often specifically mentioning an affected body part. When talking about an adverse drug reaction, tweets often mention affected organs. To identify these mentions of anatomy, we extracted a gazetteer list from sub-tree A of MeSH.

ADR We use the adverse drug reaction (ADR) lexicon provided by (Nikfarjam et al., 2015) which is a collection of several lexica including SIDER (Kuhn et al., 2016), CHV (Zeng et al., 2007), COSTART,[4] and DIEGO_Lab ADR lexicon[5].

Preg For tweets that mention a pregnancy issue, rather than a birth defect, a gazetteer list of pregnancy complication terms was extracted from subtree C13.703 of MeSH.

(13) *After a stillbirth in 2014 for #Trisomy18, yesterday we found out we are expecting a healthy baby*

BirthDef Terms referring to birth defects vary. For instance, *Down's Syndrome* is variously referred to as *Mongolism*, *Trisomy G*, and *Trisomy 21*. Supplementing the gazetteer list potentially enables the model to make correct predictions for those instances of birth defects that have not been observed in the training data. We compiled a gazetteer list of congenital, hereditary, and neonatal diseases and abnormalities from MeSH C16.

Descendant Terms referring to children: *kid, son, daughter, baby, child, fetus, girl, boy, infant, toddler, twins*. In addition, this gazetteer includes patterns such as *one year old* described by the regular expression:

CD (-|)?(day|days|month|months|year|years)(-|)?old

where ? denotes zero or one instance of a token, | denotes alternatives, and *CD* indicates a number.

FamilyRel Family terms such as *mom, dad, mommy, daddy, mother, father, grandfather, grandmother, wife, husband, spouse, sister, brother, parent, sister in law, brother in law, cousin, niece, nephew*

Acquaintances Terms such as *neighbor, friend, colleague*, etc., from GI (Stone et al., 1966) (tagged as SocRel)

The last two gazeteers enable the system to distinguish examples of self reports (*my child*) from reports on a family member (*my cousin's kid*).

[4] http://bioportal.bioontology.org/ontologies/COSTART
[5] http://diego.asu.edu/Publications/ADRMine.html

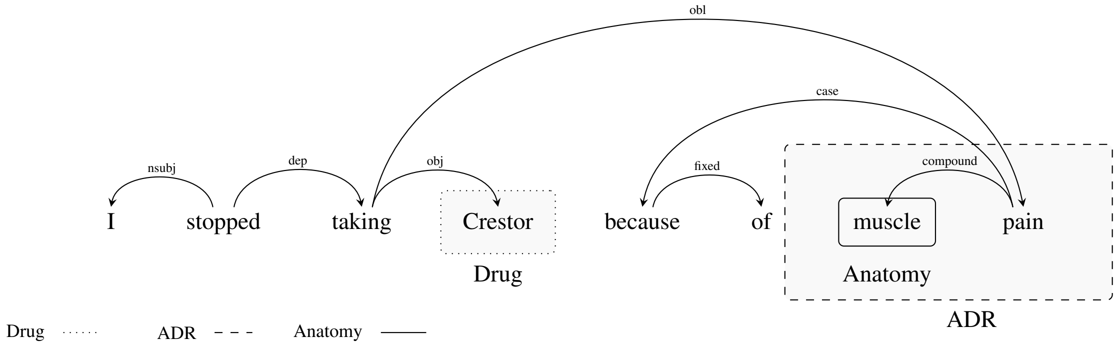

Figure 1: Dependency parse for *I stopped taking Crestor because of muscle Pain*, with gazetteer annotations

4.2 Named entities

Proper names and names of organizations are extracted using the ANNIE module.

Names People often mention the name of their child when talking about a personal experience.

(14) *My Kristin such a blessing from GOD - Kids with Down Syndrome*

Organization The presence of an organization mention often indicates that a tweet is talking in a general sense and not relating a personal experience.

(15) *Ohio Senate Says 'No' to Abortion Based on Down Syndrome Diagnosis*

Gazetteer lists have two main advantages. First, they enable the model to extend its vocabulary. Second, they determine the position of a certain annotation type in a sentence, which becomes important when coupled with dependency relations, as illustrated in Figure 1. The mapping[6] of gazetteer lists to tasks is provided in Table 2.

Table 2: The set of gazetteer lists used for each task, subsumed under the label *Gaz*

Task	Gazetteer set
SM18-2	Drug, Disease
SM18-4	Descendant, FamilyRel, Acquaintance
SM19-1	Drug, Disease, ADR, Anatomy, Descendant, FamilyRel, Acquaintance
SM20-5	BirthDef, Preg, Anatomy, Descendant, FamilyRel

4.3 POS tags

Part-of-speech tags are the most widely used linguistic feature and are available from many standard NLP environments. POS tags provide useful information such as types of pronouns and tense for verbs, all important clues for the detection of a personal experience. POS tags have been used in the literature for classifying personal and impersonal sentences (Li et al., 2010).

Since our tasks focus intensely on first person reports, we replace the single tag for pronouns *PRP* with three tags *PRP1, PRP2, PRP3*, reserved for first, second, and third person pronouns. Likewise, the reflexive pronoun tag *PRP$* is replaced by *PRP$1, PRP$2, and PRP$3* for first, second, and third person possessive pronouns. In our experiments we compare the standard Penn Treebank tag set (denoted by POS1) to this extended POS tag set (denoted by POS2).

While POS-tag information is partially encoded in word embeddings, our ablation shows that explicit encoding leads to performance increase and that POS2 is part of our best performing model.

4.4 Dependency parse

Dependency graphs provide syntactic knowledge as well as shallow semantic information. An example of a dependency graph for the ADR task together with gazetteer annotations is provided in Figure 1.

Some dependency relations are indicative of personal experience mentions. For instance, drug or birth defect mentions occur more likely as direct objects. Self-reports mostly use first person pronouns in subject position.

We use the Stanford parser (Klein and Manning, 2003) to determine dependency relations.

[6] Note that in Table 4, the label *Gaz* refers to the respective gazetteer set

	h_1^1	h_2^1	h_3^1	h_4^1	h_5^1	h_6^1	h_7^1	h_8^1	h_9^1
	=	=	=	=	=	=	=	=	=
POS Embedding	P_{NNP}	P_{VBZ}	P_{NNP}	P_{VBG}	P_{PRP3}	P_{TO}	P_{VB}	P_{NN}	P_{NN}
	+	+	+	+	+	+	+	+	+
Gazetteer Embedding	G_{Name}	**0**	$G_{BirthDef}$	**0**	**0**	**0**	**0**	$G_{Anatomy}$	**0**
	+	+	+	+	+	+	+	+	+
Token Embedding	E_{Milo}	E_{has}	$E_{Hydrocephalus}$	$E_{causing}$	E_{him}	E_{to}	E_{need}	E_{brain}	E_{shunt}
Input	Milo	has	Hydrocephalus	causing	him	to	need	brain	shunt
time-step	1	2	3	4	5	6	7	8	9

Figure 2: Additive annotation embedding (Layer1)

5 Proposed model

We developed a multi-layer system which includes four layers, namely: embeddings, self-attention, GCNN, and classification.

Layer1: Embeddings We combine traditional word embeddings with POS embeddings and our gazetteer embeddings additively.

- Tokens are embedded by GLoVE,[7] ELMo,[8], pretrained RoBERTa[9] (Liu et al., 2019), or BioBERT (Lee et al., 2020).
- We pretrain POS embeddings using Word2Vec. Our approach is to apply Word2Vec on POS tags instead of tokens. The embeddings are trained using the Gensim package (Rehurek and Sojka, 2010) with a window size of $w = 5$. The pretraining is performed on training data of all task introduced in Section 3. The resulting embeddings are used to initialize a POS embedding matrix $P \in \mathbb{R}^{\phi \times d_{emb}}$, where ϕ is the number of distinct POS tags, and d_{emb} is the dimensionality of the word embeddings. The POS embeddings are fine-tuned during the training for the main classification task.
- A gazetteer annotation x is embedded through a vector $G_x \in \mathbb{R}^{d_{emb}}$ which is a learnable parameter and is updated during the training of the main classification task.
- Inspired by BERT (Devlin et al., 2019) segment embeddings, we add POS embeddings and gazetteer embeddings to token embeddings. We call this scheme, additive annotations. Figure 2 shows that for time step 3, the vectors $E_{Hydrocephalus}$, $G_{BirthDef}$, and P_{NNP} are added to form h_3^1, the aggregate for time-step 3 in layer 1.

The additive approach enables the model to encode as many as gazetteer annotations, without introducing new dimensions to the token representations (in contrast to concatenative approaches). After the training, one can easily introduce a new gazetteer annotation y by adding a learnable vector G_y to the model parameters, and only fine-tune the model, without making any changes to hidden dimensions.

Layer2: Self-attention encoder We use a self-attention encoder (the encoder part of the Transformer) as first layer (Vaswani et al., 2017). The encoder at layer 2 gets the representations h_i^1 and outputs representations h_i^2. The number of heads in the multi-head attention is $n_{heads} = 4$ and the dimensionality of the feed-forward layer is $d_{FF} = 1024$.

Layer3: Graph CNN We use a graph convolutional network (GCNN) (Kipf and Welling, 2017) to encode the dependency graph following (Marcheggiani and Titov, 2017). In GCCN, each token is represented based on its adjacent tokens in a dependency parse using:

$$h_i^3 = ReLU(\sum_{j \in \mathcal{N}(i)} W_{L(i,j)} h_j^2 + b) \qquad (1)$$

where $\mathcal{N}(i)$ is the set of tokens adjacent to token i and $L(i,j)$ is the label of the arc from token j to token i. Note that the network is not tied, i.e. $W_{L(i,j)}$ depends on the arc labels. GCNN receives h_i^2 and outputs token-wise representations h_i^3.

Layer4: Pooling and classification For the vector representation of the tweet, attention (Bahdanau et al., 2015) is calculated from importance scores:

$$e_i = w_{att}^T h_i^3 \qquad (2)$$

[7] Twitter ($d_{emb} = 200$)
[8] Original ($d_{emb} = 1024$)
[9] Large ($d_{emb} = 1024$)

Table 3: The set of hyper-parameters used for each task

	Embedding	Epoch	lr
SM18-2	GLoVE	5	.1e-3
	ELMo	6	.5e-4
	RoBERTa / BioBERT	10	.5e-5
SM18-4	GLoVE	4	.1e-3
	ELMo	6	.1e-3
	RoBERTa / BioBERT	6	.1e-4
SM19-1	GLoVE	6	.1e-3
	ELMo	8	.1e-4
	RoBERTa / BioBERT	10	.1e-5
SM20-5	GLoVE	4	.1e-4
	ELMo	6	.1e-4
	RoBERTa / BioBERT	8	.5e-5

using a latent context vector w_{att}, and then normalizing the scores using softmax:

$$\alpha_i = \frac{exp(e_i)}{\sum_j exp(e_j)} \tag{3}$$

The normalized scores are then used for a weighted sum $H = \sum_i \alpha_i * h_i^3$. The final vector H is used as input to a linear layer for the classification.

The proposed model is implemented using the PyTorch library (Paszke et al., 2017). Cross-entropy is used to calculate the network loss and the model is optimized using the Adam optimizer (Kingma and Ba, 2015). Table 3 details the hyper-parameters used for each task.

6 Numerical results

We evaluate the proposed model using a set of ablation studies. The SM19-1 and SM20-5 tasks are evaluated on the official test data. For SM18-2 and SM18-4, official test data is not available, therefore we replicate the state-of-the art systems and perform evaluation on a hold-out set from the original training data.

Table 4 shows that all tasks benefit moderately from POS features with the extended POS tagset POS2 outperforming the standard Penn Treebank tagset POS1. POS features increase performance for GLoVE and ELMO more than for RoBERTa or BioBERT. Dependency information Dep, similarly, yields consistent small improvements. Note, however, the asymmetrically stronger improvements in precision, especially for RoBERTa and BioBERT models. Combining POS and Dep results in another consistent small improvement, showing that the features effectively interoperate.

The gazetteer lists and named-entity features provide considerable improvements for all tasks except for SM18-4 with marginal improvements. Note that SM18-4 is the vaccination behaviour prediction task, specific to flu. This result is to be expected: it requires identifying self reports, but the trigger terms for the flu domain consisted only of *flu*, making gazetteers ineffective. The tasks with more diverse vocabularies show greater impact of gazetteer lists.

Combining grammatical and gazetteer features robustly yields best results. Interestingly, adding knowledge resources to a lighter language model approaches performance of a larger model. For instance, GloVe with all resources outperforms ELMo without resources for all tasks and even approaches RoBERTa or BioBERT without resources.

We also observe that while our system configurations using RoBERTa or BioBERT reported in Table 4 beat the SOTA reported in competition in F1 and precision, our recall only exceeds SOTA for SM18-4 and SM18-4. We interpret this as a strong point of our system: in health-related applications, precision often outweighs recall. For system development, increasing recall is usually easier and this paper limits itself to gazetteer lists and linguistic features for domain adaptation to show their potential and limits. This leaves room for further error-driven domain adaptation.

7 Case-study

To examine the effects of knowledge sources, we probe the attention importance scores at Layer4 (Equation 2). The scores demonstrate how the the model attends to different tokens. We probe the scores in two cases, with and without gazetteer lists. Figure 3 and Figure 4 demonstrate visualizations of the attention scores for two samples from SM20-5 and SM19-1 tasks respectively. Higher attention scores are indicated with darker gray color.

The model of Figure 3a uses no gazetteer lists. The model partially attends to the birth defect mention *Trisomy18* and pays no attention to the pregnancy issue *StillBirth*. The model, however, properly attends to the personal pronouns *I* and lexical triggers such as *baby* and *birth*. On the other hand, when the model is given BirthDef and Pregnancy gazetteers, the model puts more attention on *StillBirth* as evidence for a birth defect, leading to a more certain prediction.

A similar pattern is observed in Figure 4. Supply-

Table 4: Ablation of grammatical features and gazetteer lists

		SM18-2			*SM18-4*			*SM19-1*			*SM20-5*		
	Features	μP	μR	μF	P	R	F	P	R	F	μP	μR	μF
GLoVE	None	.63	.64	.63	.78	.76	.77	.50	.51	.50	.54	.53	.54
	POS1	.64	.67	.66	.79	.78	.78	.51	.55	.53	.53	.57	.55
	POS2	.66	.67	.66	.79	.81	.80	.53	.55	.54	.53	.59	.56
	Dep	.68	.65	.67	.83	.75	.79	.56	.53	.54	.58	.58	.58
	POS1, Dep	.70	.66	.68	.81	.78	.79	.57	.55	.56	.60	.59	.60
	POS2, Dep	.70	.68	.69	.82	.78	.80	.58	.58	.58	.61	.60	.60
	Gaz, Name, Org	.69	.72	.71	.79	.77	.78	.55	.56	.55	.60	.58	.59
	Gaz, Name, Org, Pos2, Dep	.73	.71	.72	.83	.78	.81	.59	.60	.60	.65	.60	.62
ELMo	None	.71	.69	.70	.80	.78	.79	.54	.53	.54	.64	.60	.62
	POS1	.73	.68	.70	.79	.81	.80	.54	.54	.54	.66	.61	.63
	POS2	.72	.70	.71	.80	.81	.81	.53	.57	.55	.65	.63	.64
	Dep	.75	.68	.71	.86	.78	.82	.58	.56	.57	.68	.61	.65
	POS1, Dep	.74	.71	.72	.85	.80	.83	.57	.59	.58	.70	.62	.66
	POS2, Dep	.75	.71	.73	.86	.81	.84	.59	.60	.60	.71	.62	.67
	Gaz, Name, Org	.72	.76	.74	.80	.80	.80	.58	.56	.57	.69	.66	.68
	Gaz, Name, Org, Pos2, Dep	.77	.73	.75	.85	.83	.84	.61	.61	.61	.71	.65	.69
RoBERTa	None	.71	.74	.73	.87	.82	.85	.62	.58	.60	.68	.62	.65
	POS1	.69	.76	.73	.85	.86	.85	.61	.60	.60	.72	.64	.69
	POS2	.71	.75	.74	.87	.87	.87	.62	.60	.61	.74	.65	.70
	Dep	.76	.73	.74	.89	.86	.87	.64	.59	.61	.74	.60	.67
	POS1, Dep	.75	.75	.75	.87	.88	.87	.62	.62	.62	.75	.65	.70
	POS2, Dep	.76	.77	.76	.88	.88	.88	.63	.62	.62	.76	.65	.71
	Gaz, Name, Org	.73	.76	.75	.89	.83	.86	.65	.60	.63	.76	.65	.71
	Gaz, Name, Org, Pos2, Dep	.77	.78	.77	.90	.89	.89	.69	.62	.66	.79	.67	.73
BioBERT	None	.72	.76	.74	.85	.83	.84	.61	.63	.62	.67	.65	.66
	POS1	.73	.75	.74	.85	.86	.85	.63	.63	.63	.69	.67	.68
	POS2	.73	.75	.74	.86	.87	.86	.64	.63	.63	.69	.68	.69
	Dep	.78	.72	.75	.88	.88	.88	.66	.63	.65	.74	.62	.68
	POS1, Dep	.78	.73	.75	.88	.89	.88	.65	.65	.65	.75	.67	.71
	POS2, Dep	.78	.74	.76	.88	.89	.89	.65	.66	.65	.76	.68	.72
	Gaz, Name, Org	.74	.77	.76	.88	.84	.86	.64	.62	.63	.76	.68	.72
	Gaz, Name, Org, Pos2, Dep	.77	.77	.77	.90	.89	.89	.68	.65	.67	.79	.68	.73
	SOTA:	.63	.77	.70	.82	.79	.80	.60	.68	.64	.65	.73	.69
		(Xherija, 2018)			(Joshi et al., 2018)			(Chen et al., 2019)			(Bai and Zhou, 2020)		

ing the model with the Anatomy and Drug gazetteer leads the model to pay more attention to drug mentions and mentions of affected body parts.

8 Conclusions

This paper demonstrates the effectiveness of using gazetteer lists from high-quality sources, standard named entity categories and part-of-speech embeddings with a self-attention encoder and a GCNN encoding grammatical dependencies. The architecture supports precision oriented domain adaptation from widely available, high-quality resources (i.e. MeSH). Adaptation with new gazetteer lists using additive annotation sidesteps the need to reconfigure or retrain the neural networks

The experiments confirm that quality external resources can offset the lower parameter space of light-weight word embedding/language models, such as GLoVE and ELMo. At the same time, these resources effectively combine with RoBERTa for best performance. The stronger improvements in precision are especially promising for health applications.

The comparative results on different tasks and different domains proves that this extensible architecture is well-suited for actual use in the wild on domains and tasks, for which experts know to supply high-quality terminology resources.

Acknowledgement

We would like to thank the anonymous reviewers for their valuable and constructive comments.

Its been a year since I found out I 'd be giving birth to a sleeping baby ! # StillBirth ! #Loss # Trisomy18

a) Without Birth Defect and Pregnancy gazetteers

Its been a year since I found out I 'd be giving birth to a sleeping baby ! # StillBirth #Loss # Trisomy18

b) With Birth Defect and Pregnancy gazetteers

Figure 3: Visualization of attention scores for a sample from SM20-5

after taking olanzapine i wake up and feel like i am in a straight jacket because my muscles feel stiff

a) Without Drug and Anatomy gazetteers

after taking olanzapine i wake up and feel like i am in a straight jacket because my muscles feel stiff

b) With Drug and Anatomy gazetteers

Figure 4: Visualization of attention scores for a sample from SM19-1

References

Olanrewaju Tahir Aduragba, Jialin Yu, Gautham Senthilnathan, and Alexandra Crsitea. 2020. Sentence Contextual Encoder with BERT and BiLSTM for Automatic Classification with imbalanced medication tweets. In *SMM4H 2020.*

Segun Taofeek Aroyehun and Alexander Gelbukh. 2019. Detection of Adverse Drug Reaction in Tweets Using a Combination of Heterogeneous Word Embeddings. In *SMM4H 2019.*

Yandrapati Prakash Babu and Rajagopal Eswari. 2020. Identification of Medication Tweets Using Domain-specific Pre-trained Language Models. In *SMM4H 2020.*

Parsa Bagherzadeh, Nadia Sheikh, and Sabine Bergler. 2018. CLaC at SMM4H task 1, 2, and 4. In *SMM4H 2018.*

Parsa Bagherzadeh, Nadia Sheikh, and Sabine Bergler. 2019. Adverse Drug Effect and Personalized Health Mentions, CLaC at SMM4H 2019, Tasks 1 and 4. In *SMM4H 2019.*

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the 3rd International Conference on Learning Representations, ICLR'15.*

Yang Bai and Xiaobing Zhou. 2020. Automatic Detecting for Health-related Twitter Data with BioBERT. In *SMM4H 2020.*

Paul Barry and Ozlem Uzuner. 2019. Deep Learning for identification of adverse effect mentions in Twitter data. In *SMM4H 2019.*

Brian Chan, Andrea Lopez, and Urmimala Sarkar. 2015. The canary in the coal mine tweets: social media reveals public perceptions of non-medical use of opioids. *PloS one*, 10(8).

Lauren E Charles-Smith, Tera L Reynolds, Mark A Cameron, Mike Conway, Eric HY Lau, Jennifer M Olsen, Julie A Pavlin, Mika Shigematsu, Laura C Streichert, Katie J Suda, et al. 2015. Using social media for actionable disease surveillance and outbreak management: a systematic literature review. *PloS one*, 10(10).

Shuai Chen, Yuanhang Huang, Xiaowei Huang, Haoming Qin, Jun Yan, and Buzhou Tang. 2019. HITSZ-ICRC: A report for SMM4H shared task 2019-automatic classification and extraction of adverse effect mentions in tweets. In *SMM4H 2019.*

Çağrı Çöltekin and Taraka Rama. 2018. Drug-use Identification from Tweets with Word and Character N-grams. In *SMM4H 2018.*

Javier Cortes-Tejada, Juan Martinez-Romo, and Lourdes Araujo. 2019. NLP@ UNED at SMM4H 2019: Neural Networks Applied to Automatic Classifications of Adverse Effects Mentions in Tweets. In *SMM4H 2019.*

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina N. Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT 2019.*

Su Golder, Gill Norman, and Yoon K Loke. 2015. Systematic review on the prevalence, frequency and comparative value of adverse events data in social media. *British journal of clinical pharmacology*, 80(4).

Jeremy Howard and Sebastian Ruder. 2018. Universal Language Model Fine-tuning for Text Classification. In *ACL 2108.*

Keyuan Jiang, Ricardo Calix, and Matrika Gupta. 2016. Construction of a personal experience tweet corpus for health surveillance. In *BioNLP.*

Keyuan Jiang, Shichao Feng, Ricardo A Calix, and Gordon R Bernard. 2019. Assessment of word embedding techniques for identification of personal experience tweets pertaining to medication uses. In *International Workshop on Health Intelligence.*

Keyuan Jiang, Shichao Feng, Qunhao Song, Ricardo A Calix, Matrika Gupta, and Gordon R Bernard. 2018. Identifying tweets of personal health experience through word embedding and LSTM neural network. *BMC Bioinformatics*, 19(8).

Aditya Joshi, Xiang Dai, Sarvnaz Karimi, Ross Sparks, Cecile Paris, and C Raina MacIntyre. 2018. Shot or not: Comparison of NLP approaches for vaccination behaviour detection. In *SMM4H 2018*.

Pervaiz Iqbal Khan, Imran Razzak, Andreas Dengel, and Sheraz Ahmed. 2020. Improving Personal Health Mention Detection on Twitter Using Permutation Based Word Representation Learning. In *NeurIPS 2020*. Springer.

Diederik P Kingma and Jimmy Lei Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations, ICLR'15*.

Thomas. N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *ICLR'17*.

Ari Z. Klein, Ivan Flores, Arjun Magge, Anne-Lyse Minard, Karen O'Connor, Abeed Sarker, Elena Tutubalina, Davy Weissenbacher, and Graciela Gonzalez-Hernandez. 2020. Overview of the Fifth Social Media Mining for Health applications (SMM4H) shared tasks at COLING 2020. In *SMM4H 2020*.

Dan Klein and Christopher D. Manning. 2003. Accurate Unlexicalized Parsing. In *ACL 2003*.

Michael Kuhn, Ivica Letunic, Lars Juhl Jensen, and Peer Bork. 2016. The SIDER database of drugs and side effects. *Nucleic acids research*, 44(D1).

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Shoushan Li, Chu-Ren Huang, Guodong Zhou, and Sophia Yat Mei Lee. 2010. Employing personal/impersonal views in supervised and semi-supervised sentiment classification. In *ACL 2010*.

Carolyn E Lipscomb. 2000. Medical subject headings (MeSH). *Bulletin of the Medical Library Association*, 88(3).

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Diego Marcheggiani and Ivan Titov. 2017. Encoding sentences with graph convolutional networks for semantic role labeling. In *EMNLP 2017*.

Zulfat Miftahutdinov, Ilseyar Alimova, and Elena Tutubalina. 2019. KFU NLP team at SMM4H 2019 tasks: Want to extract adverse drugs reactions from tweets? BERT to the rescue. In *SMM4H 2019*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*.

Mark Myslín, Shu-Hong Zhu, Wendy Chapman, and Mike Conway. 2013. Using Twitter to examine smoking behavior and perceptions of emerging tobacco products. *Journal of medical Internet research*, 15(8).

Azadeh Nikfarjam, Abeed Sarker, Karen O'Connor, R Ginn Rachel, and Garciela Gonzalez. 2015. Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. *Journal of the American Medical Informatics Association*, 22(3).

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in PyTorch. In *NIPS 2017*.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global Vectors for Word Representation. In *EMNLP 2014*.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of NAACL*.

Radim Rehurek and Petr Sojka. 2010. Software framework for topic modelling with large corpora. In *In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*.

Dimitrios Rousidis, Paraskevas Koukaras, and Christos Tjortjis. 2020. Social media prediction: a literature review. *Multimedia Tools and Applications*, 79(9).

Rupsa Saha, Abir Naskar, Tirthankar Dasgupta, and Lipika Dey. 2018. Leveraging web based evidence gathering for drug information identification from tweets. In *SMM4H 2018*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. In *NeurIPS 2019*.

Sarah Sarabadani. 2019. Detection of adverse drug reaction mentions in tweets using ELMo. In *SMM4H 2019*.

Philip J Stone, Dexter C Dunphy, and Marshall S Smith. 1966. *The General Inquirer: A computer approach to content analysis*. MIT press.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS 2017*.

VG Vinod Vydiswaran, Grace Ganzel, Bryan Romas, Deahan Yu, Amy Austin, Neha Bhomia, Socheatha Chan, Stephanie Hall, Van Le, Aaron Miller, et al. 2019. Towards text processing pipelines to identify adverse drug events-related tweets: University of Michigan @ SMM4H 2019 Task 1. In *SMM4H 2019*.

Chen-Kai Wang, Hong-Jie Dai, and Bo-Hung Wang. 2019. BIGODM System in the Social Media Mining for Health Applications Shared Task 2019. In *SMM4H 2019*.

Davy Weissenbacher, Abeed Sarker, Arjun Magge, Ashlynn Daughton, Karen O'Connor, Michael J. Paul, and Graciela Gonzalez-Hernandez. 2019. Overview of the fourth Social Media Mining for Health (SMM4H) shared tasks at ACL 2019. In *SMM4H 2019*.

Davy Weissenbacher, Abeed Sarker, Michael J. Paul, and Graciela Gonzalez-Hernandez. 2018. Overview of the third Social Media Mining for Health (SMM4H) Shared Tasks at EMNLP 2018. In *SMM4H 2018*.

David S Wishart, Yannick D Feunang, An C Guo, Elvis J Lo, Ana Marcu, Jason R Grant, Tanvir Sajed, Daniel Johnson, Carin Li, Zinat Sayeeda, et al. 2018. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic acids research*, 46(D1).

Chuhan Wu, Fangzhao Wu, Zhigang Yuan, Junxin Liu, Yongfeng Huang, and Xing Xie. 2019. MSA: Jointly detecting drug name and adverse drug reaction mentioning tweets with multi-head self-attention. In *Twelfth ACM International Conference on Web Search and Data Mining*.

Orest Xherija. 2018. Classification of medication-related tweets using stacked bidirectional LSTMs with context-aware attention. In *SMM4H 2018*.

Qing Zeng, Tony Tse, Guy Divita, Alla Keselman, Jonathan Crowell, Allen Browne, Sergey Goryachev, and Long Ngo. 2007. Term identification methods for consumer health vocabulary development. *Journal of medical Internet research*, 9(1).