People on Drugs: Credibility of User Statements in Health Communities

Subhabrata Mukherjee[†] Gerhard Weikum[†] Cristian Danescu-Niculescu-Mizil[‡]
[†]Max Planck Institute for Informatics, [‡]Max Planck Institute for Software Systems
smukherjee@mpi-inf.mpg.de, weikum@mpi-inf.mpg.de, cristian@mpi-sws.org

Abstract

Online health communities are a valuable source of information for patients and physicians. However, such user-generated resources are often plagued by inaccuracies and misinformation. In this work we propose a method for automatically establishing the credibility of user-generated medical statements and the trustworthiness of their authors by exploiting linguistic cues and distant supervision from expert sources. To this end we introduce a probabilistic graphical model that jointly learns user trustworthiness, statement credibility, and language objectivity.

We apply this methodology to the task of extracting rare or unknown side-effects of medical drugs—this being one of the problems where large scale non-expert data has the potential to complement expert medical knowledge. We show that our method can reliably extract side-effects and filter out false statements, while identifying trustworthy users that are likely to contribute valuable medical information.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval - *Information Filtering*; I.2.7 [**Computing Methodologies**]: Artificial Intelligence - *Natural Language Processing*

General Terms

Design, Algorithms, Measurement, Experimentation

Keywords

Credibility; Trustworthiness; Objectivity; Veracity; Probabilistic Graphical Models

1. INTRODUCTION

Online social media includes a wealth of topic-specific communities and discussion forums about politics, music, health, and many other domains. User-generated content in such communities offer a great potential for distilling and analyzing facts and opinions. In particular, online health communities constitute an important source of information for patients and doctors alike, with 59% of the adult U. S. population consulting online health resources [13], and nearly half of U. S. physicians relying on online resources for professional use [17].

One of the major hurdles preventing the full exploitation of information from online health communities is the widespread concern regarding the quality and credibility of user-generated content [37, 48]. To address this issue, this work proposes a model that can automatically assess the credibility of medical statements made by users of online health communities. In particular, we focus on extracting rare or unknown side-effects of drugs—this being one of the problems where large scale non-expert data has the potential to complement expert medical knowledge [47], but where misinformation can have hazardous consequences [7].

The main intuition behind the proposed model is that there is an important interaction between the credibility of a statement, the trustworthiness of the user making that statement and the language used in the post containing that statement. Therefore, we consider the mutual interaction between the following factors:

- *Users:* the overall *trustworthiness* (or authority) of a user, corresponding to her status and engagement in the community.
- *Language:* the *objectivity*, rationality (as opposed to emotionality), and general quality of the language in the users' posts. Objectivity is the quality of the post to be free from preference, emotion, bias and prejudice of the author.
- *Statements:* the *credibility* (or truthfulness) of medical statements contained within the posts. Identifying accurate drug side-effect statements is a goal of the model.

These factors have a strong influence on each other. Intuitively, a statement is more credible if it is posted by a trustworthy user and expressed using confident and objective language. As an example, consider the following review about the drug Depo-Provera by a senior member of `healthboards.com`, one of the largest online health communities:
"... Depo is very dangerous as a birth control and has too many long term side-effects like reducing bone density ..."
This post contains a credible statement that a side-effect of Depo-Provera is to reduce bone density. Conversely, highly subjective and emotional language suggests lower credibility of the user's statements. A negative example along these lines is:

KDD'14, August 24–27, 2014, New York, NY, USA.

http://dx.doi.org/10.1145/2623330.2623714.

"I have been on the same cocktail of meds (10 mgs. Elavil at bedtime/60-90 mgs. of Oxycodone during the day/1/1/2 mgs. Xanax a day....once in a while I have really bad hallucination type dreams. I can actually "feel" someone pulling me of the bed and throwing me around. I know this sounds crazy but at the time it fels somewhat demonic."

Although this post suggests that taking Xanax can lead to hallucination, the style in which it is written renders the credibility of this statement doubtful. These examples support the intuition that to identify credible medical statements, we also need to assess the trustworthiness of users and the objectivity of their language. In this work we leverage this intuition through a *joint analysis of statements, users, and language* in online health communities.

Although information extraction methods using probabilistic graphical models [39, 21] have been previously employed to extract statements from user generated content, they do not account for the inherent bias, subjectivity and misinformation prevalent in health forums. Unlike standard information extraction techniques [23, 5, 41], our method considers the role language can have in assessing the credibility of the extracted statements. Stylistic features—such as the use of modals and inferential conjunctions—help identify accurate statements, while affective features help determine the emotional state of the user making those statements (e.g., anxiety, confidence).

The main technical contribution of this paper is a probabilistic graphical model which is tailored to the problem setting as to facilitate joint inference over users, language, and statements. We devise a Markov Random Field (MRF) with individual users, posts, and statements as nodes, as summarized in Figure 1. The quality of these nodes—trustworthiness, objectivity, and credibility—is modeled as binary random variables. The model is semi-supervised with a subset of training side-effect statements derived from expert medical databases, labeled as true or false. In addition, the model relies on linguistic and user features that can be directly observed in online communities. Inference and parameter estimation is done via an EM (Expectation-Maximization) framework, where MCMC sampling is used in the *E-step* for estimating the label of unknown statements and the Trust Region Newton method [27] is used in the *M-step* to compute feature weights.

We apply our method to 2.8 million posts contributed by 15,000 users of one of the largest online health community `healthboards.com`. Our model achieves an overall accuracy of 82% in identifying drug side-effects, bringing an improvement of 13% over an SVM baseline using the same features and an improvement of 4% over a stronger SVM classifier which uses distant supervision to account for feature sparsity. We further evaluate how the proposed model performs in two realistic use cases: discovering rare side-effects of drugs and identifying trustworthy users in a community.

To summarize, this paper brings the following main contributions:

- *Model:* It proposes a model that captures the interactions between user trustworthiness, language objectivity, and statement credibility in social media (Section 2), and devises a comprehensive feature set to this end (Section 3);
- *Method:* It introduces a method for joint inference over users, language, and statements (Section 4) through a probabilistic graphical model;
- *Application:* It applies this methodology to the problem of extracting side-effects of medical drugs from online health forums (Section 5);
- *Use-cases:* It evaluates the performance of the model in the context of two realistic practical tasks (Section 6).

2. OVERVIEW OF THE MODEL

Our approach leverages the intuition that there is an important interaction between statement credibility, linguistic objectivity, and user trustworthiness. We therefore model these factors jointly through a probabilistic graphical model, more specifically a Markov Random Field (MRF), where each statement, post and user is associated with a binary random variable. Figure 1 provides an overview of our model. For a given statement, the corresponding variable should have value 1 if the statement is credible, and 0 otherwise. Likewise, the values of post and user variables reflect the objectivity and trustworthiness of posts and users.

Nodes, Features and Labels Nodes associated with users and posts have observable features, which can be extracted from the online community. For users, we derive engagement features (number of questions and answers posted), interaction features (e.g., replies, giving thanks), and demographic information (e.g., age, gender). For posts, we extract linguistic features in the form of discourse markers and affective phrases. Our features are presented in details in Section 3. While for statements there are no observable features, we can derive distant training labels for a subset of statements from expert databases, like the Mayo Clinic,[1] which list typical as well as rare side-effects of widely used drugs.

Edges The primary goal of the proposed system is to retrieve the credibility label of unobserved statements given *some* expert labeled statements and the observed features by leveraging the mutual influence between the model's variables. To this end, the MRF's nodes are connected by the following (undirected) edges:

- each user is connected to all her posts;
- each statement is connected to all posts from which it can be extracted (by state of the art information extraction methods);
- each user is connected to statements that appear in at least one of her posts.

Configured this way, the model has the capacity to capture important interactions between statements, posts, and users — for example, credible statements can boost a user's trustworthiness, whereas some false statements may bring it down. Furthermore, since the inference (detailed in Section 4) is centered around the cliques in the graph (factors) and multiple cliques can share nodes, more complex "cross-talk" is also captured. For instance, when several highly trustworthy users agree on a statement and one user disagrees, this reduces the trustworthiness of the disagreeing user.

In addition to establishing the credibility of statements, the proposed system also computes individual likelihoods as a by-product of the inference process, and therefore can output rankings for all statements, users, and posts, in descending order of credibility, trustworthiness, and objectivity.

[1] `mayoclinic.org/drugs-supplements/`

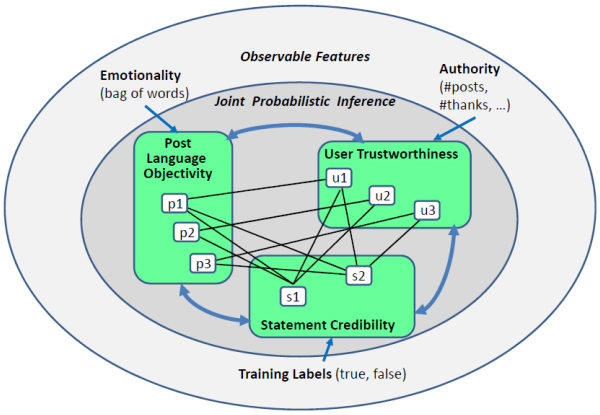


Figure 1: Overview of the proposed model, which captures the interactions between statement credibility, post objectivity, and user trustworthiness.

Feature types	Example values
Strong modals	might, could, can, would, may
Weak modals	should, ought, need, shall, will
Conditionals	if
Negation	no, not, neither, nor, never
Inferential conj.	therefore, thus, furthermore
Contrasting conj.	until, despite, in spite, though
Following conj.	but, however, otherwise, yet
Definite det.	the, this, that, those, these
First person	I, we, me, my, mine, us, our
Second person	you, your, yours
Third person	he, she, him, her, his, it, its
Question particles	why, what, when, which, who
Adjectives	correct, extreme, long, visible
Adverbs	maybe, about, probably, much
Proper nouns	Xanax, Zoloft, Depo-Provera

Table 1: Stylistic features.

3. FEATURES

3.1 Linguistic Features

The linguistic characteristics of a post can convey the author's attitude towards her statements as well as her emotional state [8]. In our model we use *stylistic* and *affective* features to assess a post's objectivity and quality.

Stylistic Features Consider the following user post:

"I heard Xanax can have pretty bad side-effects. You may have peeling of skin, and apparently some friend of mine told me you can develop ulcers in the lips also. If you take this medicine for a long time then you would probably develop a lot of other physical problems. Which of these did you experience ?"

This post evokes a lot of uncertainty, and does not specifically point to the occurrence of any side effect from a first-hand experience. Note the usage of strong modals (depicting a high degree of uncertainty) "can", "may", "would", the indefinite determiner "some", the conditional "if", the adverb of possibility "probably" and the question particle "which". Even the usage of too many named entities for drug and disease names can impact the credibility of a statement (refer the introductory example).

Contrast the above post with the following one :

"Depo is very dangerous as a birth control and has too many long term side-effects like reducing bone density. Hence, I will never recommend anyone using this as a birth control. Some women tolerate it well but those are the minority. Most women have horrible long lasting side-effects from it."

This post uses the inferential conjunction "hence" to draw conclusions from a previous argument, the definite determiners "this", "those", "the" and "most" to pinpoint entities and the highly certain weak modal "will".

Table 1 shows a set of linguistic features which we deem suitable for discriminating between these two kinds of posts. Many of the features related to epistemic modality have been discussed in prior linguistic literature [8, 46] and features related to discourse coherence have also been employed in earlier computational work (e.g., [31, 51]).

For each stylistic feature type f_i shown in Table 1 and each post p_j, we compute the relative frequency of words of type f_i occurring in p_j, thus constructing a feature vector $F^L(p_j) = \langle freq_{ij} = \#(words\ in\ f_i)\ /\ length(p_j)\rangle$. We further aggregate these vectors over all posts p_j by a user u_k into

$$F^L(u_k) = \langle \sum_{p_j\ by\ u_k} \#(words\ in\ f_i)\ /\ \sum_{p_j\ by\ u_k} length(p_j)\rangle. \quad (1)$$

Affective Features Each user has an *affective state* that depicts her attitude and emotions that are reflected in her posts. Note that a user's affective state may change over time; so it is a property of posts, not of users per se. As an example, consider the following post:

"I've had chronic depression off and on since adolescence. In the past I've taken Paxil (made me anxious) and Zoloft (caused insomnia and stomach problems, but at least I was mellow). I have been taking St. John's Wort for a few months now, and it helps, but not enough. I wake up almost every morning feeling very sad and hopeless. As afternoon approaches I start to feel better, but there's almost always at least a low level of depression throughout the day."

The high level of depression and negativity in the post makes one wonder if the statements on drug side-effects are really credible. Contrast this post to the following one:

"A diagnosis of GAD (Generalized Anxiety Disorder) is made if you suffer from excessive anxiety or worry and have at least three symptoms including...If the symptoms above, touch a chord with you, do speak to your GP. There are effective treatments for GAD, and Cognitive Behavioural Therapy in particular can help you ..."

where the user objectivity and positivity in the post make it much more credible.

We use the WordNet-Affect lexicon [40], where each word sense (WordNet synset) is mapped to one of 285 attributes of the affective feature space, like *confusion, ambiguity, hope, anticipation, hate*. We do not perform word sense disambiguation (WSD), and instead simply take the most common sense of a word (which is generally a good heuristics for WSD). For each post, we create an affective feature vector $\langle F^E(p_j)\rangle$ using these features, analogous to the stylistic vectors $\langle F^L(p_j)\rangle$. Table 2 shows a sample of the affective features used in this work.

Preliminary Feature Exploration To test whether the linguistic features introduced so far are sufficiently informative of how helpful a user is in the context of health forums, we conduct a preliminary experimental study. In

Sample Affective Features
affection, antipathy, anxiousness, approval, compunction, confidence, contentment, coolness, creeps, depression, devotion, distress, downheartedness, eagerness, edginess, embarrassment, encouragement, favor, fit, fondness, guilt, harassment, humility, hysteria, ingratitude, insecurity, jitteriness, levity, levitygaiety, malice, misery, resignation, selfesteem, stupefaction, surprise, sympathy, togetherness, triumph, weight, wonder

Table 2: Examples of affective features.

the `healthboards.com` forum, community members have the option of expressing their gratitude to a user if they find one of her posts helpful by giving "thanks" votes. Solely for the purpose of this preliminary experiment, we use the *total number of "thanks" votes* that a user received from all her posts as a weak proxy measure for user helpfulness.

We train a regression model on the per-user stylistic feature vectors $F^L(u_k)$ with #thanks normalized by #posts for each user u_k as response variable. We repeat the same experiment using only the per-user affective feature vectors $F^E(u_k)$ to identify the most important affective features.

Figure 2 shows the relative weight of various stylistic and affective linguistic features in determining user *helpfulness*, with positive weights being indicative of features contributing to a user being considered helpful by the community. Figure 2a shows that user confidence, pride, affection and positivity in statements are correlated with user helpfulness, in contrast to misery, depression and negativity in attitude. Figure 2b shows that inferential statements about definite entities have a positive impact, as opposed to the use of hypothetical statements, contrasting sentences, doubts and queries.

This experiment confirms that linguistic features can be informative in the context of online health communities. Although we use "thanks" votes as a proxy for user helpfulness, there is no guarantee that the information provided by helpful users is actually correct. A user can receive "thanks" for a multitude of reasons (e.g. being compassionate or supportive), and yet provide incorrect information. Hence, while the features described here are part of our final model, the feature weights learned in this preliminary experiment are not going to be used; instead, partially provided expert information is used to train our probabilistic model (refer to Section 4).

3.2 User Features

User demographics like age, gender and location, as well as engagement in the community reflected by the number of posts, questions, replies, or thanks received, are expected to correlate with user authority in social networks. Also, users who write long posts tend to deviate from the topic, often with highly emotional digression. On the other hand, short posts can be regarded as being crisp, objective and on topic. We attempt to capture these intuitive aspects as additional per-user features $< F^U(u_k) >$.[2]

4. PROBABILISTIC INFERENCE

As outlined in Section 2, we model our learning task as a Markov Random Field (MRF), where the random variables are the users $U = \{u_1, u_2, ...u_{|U|}\}$, their posts $P = \{p_1, p_2...p_{|P|}\}$, and the distinct statements $S = \{s_1, s_2...s_{|S|}\}$ about drug side-effects extracted from all posts. Our model is semi-supervised in that we harness ground-truth labels for a subset of statements, derived from the expert databases. Let S^L be the set of statements labeled by an expert as true or false, and let S^U be the set of unlabeled statements. Our goal is to infer labels for the statements in S^U.

The cliques in our MRF are triangles consisting of a statement s_i, a post p_j that contains that statement, and a user u_k who wrote this post. As the same statement can be made in different posts by the same or other users, there are more cliques than statements. For convenient notation, let S^* denote the set of statement instances that correspond to the set of cliques, with statements "repeated" when necessary.

Let $\phi_i(S_i^*, p_j, u_k)$ be a potential function for clique i. Each clique has a set of associated feature functions F_i with a weight vector W. We denote the individual features and their weights as f_{il} and w_l. The features are constituted by the stylistic, affective, and user features explained in Section 3: $F_i = F^L(p_j) \cup F^E(p_j) \cup F^U(u_k)$.

Instead of computing the joint probability distribution $Pr(S, P, U; W)$ like in a standard MRF, we adopt the paradigm of Conditional Random Fields (CRF's) and settle for the simpler task of estimating the conditional distribution:

$$Pr(S|P, U; W) = \frac{1}{Z(P, U)} \prod_i \phi_i(S_i^*, p_j, u_k; W), \quad (2)$$

with normalization constant $Z(P, U)$; or with features and weights made explicit:

$$Pr(S|P, U; W) = \frac{1}{Z(P, U)} \prod_i \exp(\sum_l w_l \times f_{il}(S_i^*, p_j, u_k)). \quad (3)$$

CRF parameter learning usually works on fully observed training data. However, in our setting, only a subset of the S variables have labels and we need to consider the partitioning of S into S^L and S^U:

$$Pr(S^U, S^L|P, U; W) = \frac{1}{Z(P, U)} \prod_i \exp(\sum_l w_l \times f_{il}(S_i^*, p_j, u_k)). \quad (4)$$

For parameter estimation, we need to maximize the marginal log-likelihood:

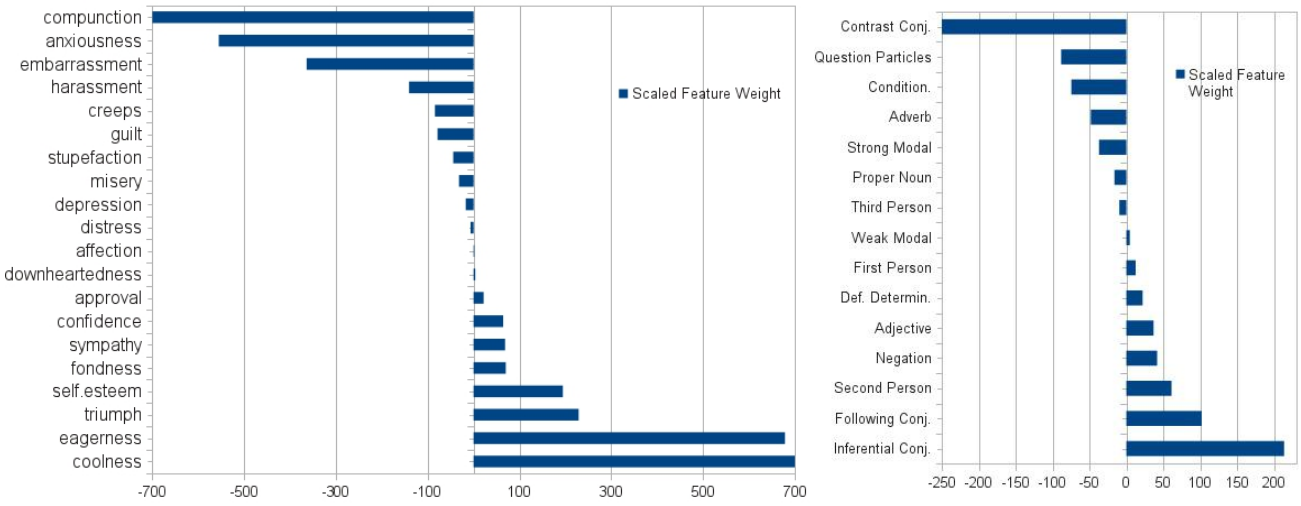
$$LL(W) = \log Pr(S^L|P, U; W) = \log \sum_{S^U} Pr(S^L, S^U|P, U; W). \quad (5)$$

We can clamp the values of S^L to their observed values in the training data [42, 54] and compute the distribution over S^U as:

$$Pr(S^U|S^L, P, U; W) = \frac{1}{Z(S^L, P, U)} \prod_i \exp(\sum_l w_l \times f_{il}(S_i^*, p_j, u_k)). \quad (6)$$

There are different ways of addressing the optimization problem for finding the argmax of $LL(W)$. In this work, we choose the Expectation-Maximization (EM) approach [29]. We first estimate the labels of the variables S^U from the posterior distribution using Gibbs sampling, and then maximize

[2]For verbosity, we compute the first three moments of each user's post-length distribution (#sentences and #words).



(a) Weight of top 20 affective features.

(b) Weight of stylistic features.

Figure 2: Relative importance of linguistic features for predicting user helpfulness in a preliminary experiment.

the log-likelihood to estimate the feature weights:

$$E - Step : q(S^U) = Pr(S^U|S^L, P, U; W^{(\nu)}) \tag{7a}$$

$$M{-}Step : W^{(\nu+1)} = argmax_{W'} \sum_{S^U} q(S^U) \log Pr(S^L, S^U|P, U; W'). \tag{7b}$$

The update step to sample the labels of S^U variables by Gibbs sampling is given by:

$$Pr(S_i^U|P, U, S^L; W) \propto \prod_{\nu \in C} \phi_\nu(S_\nu^*, p_j, u_k; W), \tag{8}$$

where C denotes the set of cliques containing statement S_i^U.

For the M-step in Equation 7b, we use an L_2-regularized Trust Region Newton Method [27], suited for large-scale unconstrained optimization, where many feature values may be zero. For this we use an implementation of LibLinear [12].

The above approach captures user trustworthiness implicitly via the weights of the feature vectors. However, we may want to model user trustworthiness in a way that explicitly aggregates over all the statements made by a user. Let t_k denote the trustworthiness of user u_k, measured as the fraction of her statements that were considered true in the previous EM iteration:

$$t_k = \frac{\sum_i \mathbb{1}_{S_{i,k}=\text{True}}}{|S_k|}, \tag{9}$$

where $S_{i,k}$ is the label assigned to u_k's statement S_i in the previous EM iteration. Equation 8 can then be modified into:

$$Pr(S_i^U|P, U, S^L; W) \propto \prod_{\nu \in C} t_k \times \phi_\nu(S_\nu^*, p_j, u_k; W) \tag{10}$$

Therefore, the random variable for trustworthiness depends on the proportion of *true* statements made by the user. The *label* of a statement, in turn, is determined by the language objectivity of the posts and trustworthiness of all the users in the community that make the statement.

The inference is an iterative process consisting of the following 3 main steps:

1. Estimate user trustworthiness t_k using Equation 9.
2. Apply the E-Step to estimate $q(S^U; W^{(\nu)})$
   For each i, sample S_i^U from Equation 7a and 10.
3. Apply the M-Step to estimate $W^{(\nu+1)}$ using Equation 7b.

5. EXPERIMENTAL EVALUATION

In this section, we study the predictive power of our probabilistic model and compare it to three baselines.

5.1 Data

We use data from the `healthboards.com`, one of the largest online health communities, with 850,000 registered members and over 4.5 million posted messages. We extracted 15,000 users and all of their posts, 2.8 million posts in total. Users are sampled based on their post frequency; Table 3 shows the user categorization in terms of their community engagement.[3] We employ an IE tool [11] to extract side-effect statements from the posts. Details of the experimental setting are available on our website.[4]

As ground truth for drug side-effects, we rely on data from the Mayo Clinic portal,[5] which contains curated expert information about drugs, with side-effects being listed as *more common, less common and rare* for each drug. We extracted 2,172 drugs which are categorized into 837 drug families. For our experiments, we select 6 widely used drug families (based on `webmd.com`). Table 4 provides information on this sample and its coverage on `healthboards.com`. Table 5 shows the number of common, less common, and rare side-effects for the six drug families as given by the Mayo Clinic portal.

[3]Overall, 77.7% of the active contributors are female.
[4]`http://www.mpi-inf.mpg.de/impact/peopleondrugs/`
[5]`mayoclinic.org/drugs-supplements/`

Member Type	Members	Posts	Average Qs.	Average Replies
Administrator	1	-	363	934
Moderator	4	-	76	1276
Facilitator	16	> 4700	83	2339
Senior veteran	966	> 500	68	571
Veteran	916	> 300	41	176
Senior member	4321	> 100	24	71
Member	5846	> 50	13	28
Junior member	1423	> 40	9	18
Inactive	1433	-	-	-
Registered user	70	-	-	-

Table 3: User statistics.

Drugs	Description	Users	Posts
alprazolam, niravam, xanax	relieve symptoms of anxiety, depression, panic disorder	2785	21112
ibuprofen, advil, genpril, motrin, midol, nuprin	relieve pain, symptoms of arthritis, such as inflammation, swelling, stiffness, joint pain	5657	15573
omeprazole, prilosec	treat acidity in stomach, gastric and duodenal ulcers, ...	1061	3884
metformin, glucophage, glumetza, sulfonylurea	treat high blood sugar levels, sugar diabetes	779	3562
levothyroxine, tirosint	treat hypothyroidism: insufficient hormone production by thyroid gland	432	2393
metronidazole, flagyl	treat bacterial infections in different body parts	492	1559

Table 4: Information on sample drug families: number of posts and number of users reporting at least one side effect.

5.2 Baselines

We compare our probabilistic model against the following baseline methods, using the same features as our model and classifying the same set of side-effect candidates.

Frequency Baseline For each statement on a drug side-effect, we consider how frequently the statement has been made in community. This gives us a ranking of side-effects.

SVM Baseline For each drug and possible side-effect we determine all posts where it is mentioned and aggregate the features F^L, F^E, F^U, described in Section 3 over all these posts, thus creating a single feature vector for each side-effect.

We use the ground-truth labels from the Mayo Clinic portal to train a Support Vector Machine (SVM) classifier with a linear kernel, L_2 loss, and L_1 or L_2 regularization, for classifying unlabeled statements.

Drug family	Common	Less common	Rare
alprazolam	35	91	45
ibuprofen	30	1	94
omeprazole	-	15	20
metformin	24	37	5
levothyroxine	-	51	7
metronidazole	35	25	14

Table 5: Number of common, less common, and rare side-effects listed by experts on Mayo Clinic.

SVM Baseline with Distant Supervision As the number of common side-effects for any drug is typically small, the above approach to create a single feature vector for each side-effect results in a very small training set. Hence, we use the notion of *distant supervision* to create a rich, expanded training set.

A feature vector is created for *every mention* or instance of a side-effect in different user posts. The feature vector $< S_i, p_j, u_k >$ has the label of the side-effect, and represents the set of cliques in Equation 2. The semi-supervised CRF formulation in our approach further allows for information sharing between the cliques to estimate the labels of the unobserved statements from the expert-provided ones.

This process creates a noisy training set, as a post may contain multiple side-effects, positive and negative. This results in multiple similar feature vectors with different labels. During testing, the same side-effect may get different labels from its different instances. We take a majority voting of the labels obtained by a side-effect, across predictions over its different instances, and assign a unique label to it.

5.3 Experiments and Quality Measures

We conduct two lines of experiments, with different settings on what is considered ground-truth.

Experimental Setting I We consider only *most common side-effects* listed by the Mayo Clinic portal as positive ground-truth, whereas all other side-effects (less common, rare and unobserved) are considered to be negative instances (i.e., so unlikely that they should be considered as false statements, if reported by a user). The training set is constructed in the same way. This setting aims to study the predictive power of our model in determining the common side-effects of a drug, in comparison to the baselines.

Experimental Setting II Here we address our original motivation: discovering less common and rare side-effects. During training, as positive ground-truth we consider common and less common side-effects (as stated by the experts on the Mayo Clinic site), whereas all rare and unobserved side-effects are considered negative instances. Our goal here is to test how well the model can identify *less known* and *rare* side-effects as true statements.

We purposely do not consider rare side-effects as positive training examples, as users frequently talk about experiencing all possible side-effects. Instead we aim to evaluate the model's ability to retrieve such statements starting only from very reliable positive instances. We measure performance on rare side-effects as the recall for such statements being labeled as true statements, in spite of considering *only* common and less common side-effects as positive instances during training.

Train-Test Data Split For each drug family, we create multiple random splits of 80% training data and 20% test data. All results reported below are averaged over 200 such splits. All baselines and our CRF model use same test sets.

Evaluation Metrics The standard measure for the quality of a binary classifier is *accuracy*: $\frac{tp+tn}{tp+fn+tn+fp}$. We also report the *specificity* ($\frac{tn}{tn+fp}$) and *sensitivity* ($\frac{tp}{tp+fn}$). Sensitivity measures the true positive rate or the model's ability to identify positive side-effects, whereas specificity measures true negative rate.

Drugs	Post Freq.	SVM			CRF
		w/o DS	DS		
			L_1	L_2	
Alprazolam	57.8	70.2	73.3	73.1	79.4
Metronidazole	55.8	68.8	79.8	78.5	82.6
Omeprazole	60.6	71.1	76.8	79.2	83.2
Levothyroxine	57.5	76.8	69.0	76.3	80.5
Metformin	55.7	53.2	79.3	81.6	84.7
Ibuprofen	58.4	74.2	77.8	80.3	82.8

Table 6: Accuracy comparison in setting I.

Drugs	Sensitivity	Specificity	Rare SE Recall	Accuracy
Metformin	79.8	91.2	99.0	86.1
Levothyroxine	89.5	74.5	98.5	83.4
Omeprazole	80.8	88.8	89.5	85.9
Metronidazole	75.1	93.8	71.0	84.2
Ibuprofen	76.6	83.1	69.9	80.9
Alprazolam	94.3	68.8	61.3	74.7

Table 7: CRF performance in setting II.

5.4 Results and Discussions

Table 6 shows the accuracy comparison of our system (CRF) with the baselines for different drug families in the first setting. The first naive baseline, which simply considers the frequency of posts containing the side-effect by different users, has an average accuracy of 57.6% across different drug families.

Incorporating supervision in the classifier as the first SVM baseline (SVM w/o DS), along with a rich set of features for users, posts and language, achieves an average accuracy improvement of 11.4%. In the second SVM baseline (SVM DS), we represent each post reporting a side-effect as a separate feature vector. This not only expands the training set leading to better parameter estimation, but also represents the set of cliques in Equation 2 (we therefore consider this to be a strong baseline). This brings an average accuracy improvement of 7% when using L_1 regularization and 9% when using L_2 regularization. Our model (CRF), by further considering the coupling between users, posts and statements, allows information to flow between the cliques in a feedback loop bringing a further accuracy improvement of 4% over the strong SVM DS L_2 baseline.

Figure 3 shows the sensitivity and specificity comparison of the baselines with the CRF model. Our approach has an overall 5% increase in sensitivity and 3% increase in specificity over the SVM L_2 baseline.

The specificity increase over the SVM L_2 baseline is maximum for the Alprazolam drug family at 8.3%. Users taking such anti-depressants often suffer from anxiety disorder, panic attacks or depression and report a large number of side-effects; also there are a large number of expert-reported side-effects for this drug family (refer Table 5). Hence, the task of discarding certain side-effects is harder for this particular drug, but our linguistic features help our model overcome this and perform well.

The drugs Metronidazole, Metformin and Omeprazole treat some serious physical conditions, have less number of expert and user-reported side-effects. Consequently, our model captures user statement corroboration well to attain a sensitivity improvement of 7.8%, 6.5% and 6.3% respectively. Overall, our classifier performs best for these drug categories.

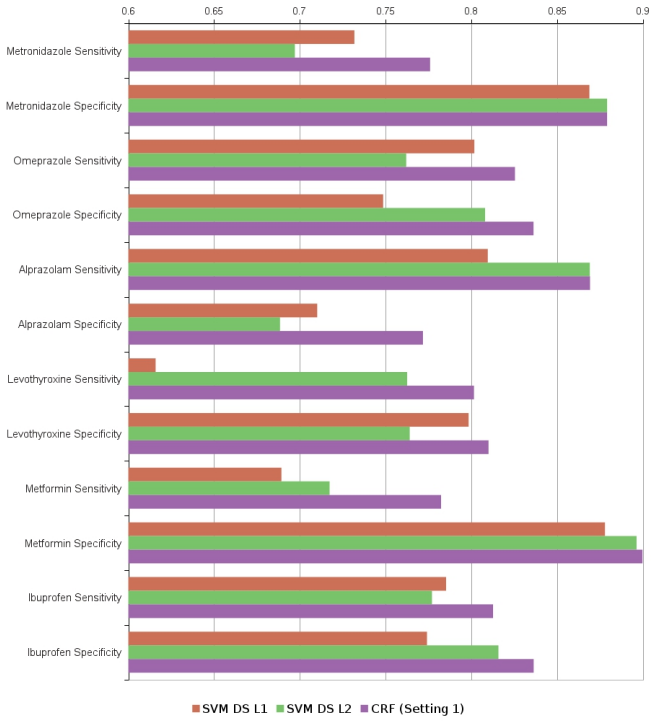


Figure 3: Specificity and sensitivity comparison of models.

Table 7 shows the overall model performance, as well as the recall for identifying rare side-effects of each drug in the second setting. The drugs Metformin, Levothyroxine and Omeprazole have much fewer side-effects, and the classifier does an almost perfect job in identifying all of them.

Feature Informativeness In order to find the *predictive power* of individual feature classes, tests are performed using L_2-loss and L_2-regularized SVM over a split of the test data. Affective features are found to be the most informative, followed by document length statistics, which are more informative than user and stylistic features. The importance of the document length features support our intuition that objective posts tend to be crisp, whereas longer ones often indulge in emotional digressions.

Among user features, the most informative is the ratio of number of replies to number of questions, followed by gender, number of posts and, finally, the number of thanks received from fellow users.

When considered independently, user, affective and stylistic features achieve F_1 scores between 51% and 55% for Alprazolam; whereas the combination of all features yield 70% F_1 score.

6. USE-CASE EXPERIMENTS

The previous section has focused on evaluating the predictive power of our model and inference method. Now we shift the focus to two application-oriented use-cases: discovering side-effects that are not covered by expert databases, and identifying the most trustworthy users that are worth following.

6.1 Discovering Rare Side Effects

Members of an online community may report side-effects that are either flagged as very rare in an expert knowledge base (KB) or not listed at all. We call the latter *out-of-KB* statements. As before, we use the data from the Mayo Clinic portal as our KB, and focus on the following two drugs representing different kinds of medical conditions and patient-reporting styles: Alprazolam and Levothyroxine. For each of these drugs, we perform an experiment as follows.

For each drug X, we identify all side-effects S that are reported for X by members of the health community; here we consider all side-effects listed for *any* drug in the KB as a potential result. For example, if "hallucination" is listed for some drug but not for the drug Xanax, we capture mentions of hallucination in posts about Xanax. We use our probabilistic model to compute credibility scores for these out-of-KB side-effects, and compile a ranked list of 10 highest-scoring side-effects for each drug. This ranked list is further extended by 10 randomly chosen out-of-KB side-effects (if reported at least once for the given drug).

The ranked list of out-of-KB side-effects is shown to two annotators[6] who manually assess their credibility, by reading the complete discussion thread (including expert replies to patient posts) and other threads that involve the users who reported the side-effect. The assessment is binary: the side-effect is considered either true (1) or false (0); we choose the final label via majority voting, breaking ties using other expert databases (`patient.co.uk` and `webmd.com`). This way, we can compute the quality of the ranked list in terms of the NDCG (Normalized Discounted Cumulative Gain) measure [18] $NDCG_p = \frac{DCG_p}{IDCG_p}$, where

$$DCG_p = rel_1 + \sum_{i=2}^{p} \frac{rel_i}{\log_2 i}. \quad (11)$$

Here, rel_i is the graded relevance of a result (0 or 1 in our case) at position i. DCG penalizes relevant items appearing lower in the rank list, where the graded relevance score is reduced logarithmically proportional to the position of the result. As the length of lists may vary for different queries, DCG scores are normalized using the ideal score, IDCG where the results of a rank list are sorted by relevance giving the maximum possible DCG score. We also report the Cohen's Kappa inter-annotator agreement measure.

Table 8 shows the Kappa and NDCG score comparison between the baseline and our CRF model. The baseline here is to rank side-effects by frequency, i.e., how often are they reported in the posts of different users on the given drug. The strength of Kappa is considered "moderate" (but significant), which depicts the difficulty in identifying the side-effects of a drug just by looking at user posts in a community. The baseline performs very poorly for the anti-depressant Alprazolam, as users suffering from anxiety disorders report a large number of side-effects most of which are not credible. On the other hand, for Levothyroxine (a drug for hypothyroidism) the baseline model performs quite well, as users report more serious symptoms and conditions associated with the drug (which also has much less expert-stated side-effects compared to Alprazolam, as shown in Table 4). The CRF model performs perfectly for both drugs.

[6]None of authors were among the annotators.

Drug	Kappa	Model NDCG Scores	
		Frequency	CRF
Alprazolam, Xanax	0.47	0.31	1
Levothyroxine, Tirosint	0.41	0.94	1

Table 8: Use-case experiment on discovering rare side-effects.

Drug	Kappa	Model NDCG Scores	
		Most-thanked	CRF
Alprazolam, Xanax	0.78	0.82	1
Levothyroxine, Tirosint	0.80	0.57	0.81

Table 9: Use-case experiment on identifying trustworthy users.

6.2 Following Trustworthy Users

In the second use-case experiment, we evaluate how well our model can identify trustworthy users in a community. We find the top-ranked users in the community given by their trustworthiness scores (t_k, as defined in Section 4), for each of the drugs Alprazolam and Levothyroxine. The baseline model selects the most-thanked contributors in the community. The moderators and facilitators of the community, listed by both models as top users, are removed from the ranked lists, in order to focus on the interesting, non-obvious cases. Two annotators are asked to annotate the top-ranked users listed by each model as trustworthy or not, based on the users' posts on the target drug. The judges are asked to mark a user as trustworthy if they would consider following the respective user in the community. Judgements were aggregated via majority voting, with ties being considered as not trustworthy. Although this task may seem highly subjective, the Cohen's Kappa scores show high inter-annotator agreement (Table 9). The strength of agreement is considered to be "very good" for the user posts on Levothyroxine, and "good" for the Alprazolam users. Also in this use-case, our model performs well and outperforms the baseline for both drug families.

7. RELATED WORK

Subject-Predicate-Object statement extraction There is ample work on extracting Subject-Predicate-Object (SPO) statements from natural-language text [39, 23, 5, 41]. State-of-the-art methods combine pattern matching with extraction rules and consistency reasoning. This can be done either in a shallow manner, over sequences of text tokens, or in combination with deep parsing and other linguistic analyses. The resulting SPO triples often have highly varying confidence, as to whether they are really expressed in the text or picked up spuriously. Judging the credibility of statements is out of the scope of classic SPO extraction methods.

Biomedical Information Extraction Customized IE techniques have been developed to tap biomedical publications like PubMed articles for extracting facts about diseases, symptoms, and drugs. Emphasis has been on the molecular level, i.e. proteins, genes, and regulatory pathways (e.g., [6, 22, 4]), and to a lesser extent on biological or medical events from scientific articles and from clinical narratives [19, 52]. LDA-style models have been used for summarizing drug-experience reports [36] and for building large knowledge bases for life science and health [11]. More recently, search engine query logs were shown to be a valuable source for identifying unknown drug side-effects [47]. Our

work is complementing these approaches, by emphasizing the role of user generated content on social media.

Truth Finding Our work relates to a research direction that aims to assess the truth of a given statement that is frequently observed on the Web—a typical example being "Obama is a Muslim" [53, 33, 35]. Information-retrieval techniques are used to systematically generate alternative hypotheses for a given statement—"Obama is a Christian"—and assess the evidence for each alternative [25]. Similar approaches have been developed for structured data such as flight times or stock quotes, where different Web sources often yield contradictory values [24]. Recently, an LDA-style latent-topic model was used for discriminating true from false claims, with various ways of generating incorrect statements (guesses, mistakes, lies) [34]. None of this prior work considered online discussion forums. Truth assessment for medical claims about diseases and their treatments (including drugs and general phrases such as "surgery") was casted as an information retrieval style evidence-aggregation and ranking method over curated health portals [45]. Although these are elaborate models, they are not geared for our setting where the credibility of statements is intertwined with user trustworthiness and the linguistic properties of user posts.

Language Analysis for Social Media Social media is an important setting for linguistic tasks that relate to our work, such as sentiment analysis (e.g., [43, 32, 28, 31, 30]), identifying bias [14, 38] and, more broadly, characterizing subjective language [49, 26]. Particularly relevant to our research direction is the link between subjectivity analysis and information extraction [50].

Trust and Reputation Management A lot of work has been dedicated to building trust and reputation management systems in social media, mostly motivated by the need to filter and organize customer product reviews, but also in the context of social networks. One type of approach has been to model the propagation of trust within a network of users [20, 15]. TrustRank [20] has become a popular measure of trustworthiness, based on random walks on (or spectral decomposition of) the user graph. Reputation management has been studied in multiple contexts, such as peer-to-peer systems, blogs, and online interactions [1, 2, 9, 3, 16]. Most of this work focused on explicit relationships between users to infer authority and trust levels, and make little or no use of the content. An exception is a model for trust propagation which devises a HITS-style algorithm for propagating trust scores in a heterogeneous network of claims, news sources, and news articles [44], building on an intuition similar to that behind our proposed approach. Evidence for a claim is collected from related news articles using generic IR-style word-level measures. In contrast, our work considers user-generated content which is represented by rich linguistic features and employs a CRF to model the complex interaction characteristic of online communities.

8. CONCLUSION

Discussions in online communities are often plagued by inaccuracies and misinformation. This hinders the exploitation of these rich and valuable resources as information sources. In this work we focus on establishing the credibility of side-effect statements in health communities. To this end, we propose a probabilistic graphical model to jointly learn the interactions between user trustworthiness, statement credibility and language use. We apply the model to extract side-effects of drugs from health communities, where we leverage the user interactions, stylistic and affective features of language use, and user properties to learn the credibility of user statements. We show that our approach is effective in reliably extracting side-effects of drugs and filtering out false information prevalent in online health communities.

In addition to validating our system's performance against expert knowledge, we show it can be successfully used in two application oriented use-cases: identifying unknown side-effects of drugs, a scenario where large-scale non-expert data has the potential to complement expert knowledge, and selecting trustworthy users that are deemed worth following.

Although our model achieves high accuracy in most of the test cases, it relies on a relatively simple information extraction machinery to identify candidate side-effect statements, which is prone to errors. The tool misses out on certain kinds of paraphrases (e.g. "nightmares" and "unusual dream" for Xanax) resulting in a drop in recall. We believe that a more sophisticated information extraction approach can further improve our approach.

Acknowledgements We thank the anonymous reviewers for their helpful (and credible) comments. We also thank Patrick Ernst for helping setting up the statement extraction tool and Amy Siu for helping in the annotation task.

9. REFERENCES

[1] B.T. Adler, L. Alfaro. A content-driven reputation system for the Wikipedia. WWW, 2007.

[2] N. Agarwal, H. Liu. Trust in Blogosphere. Encyclopedia of Database Systems, 2009.

[3] L. Alfaro, A. Kulshreshtha, I. Pye, B.T. Adler. Reputation systems for open collaboration. Commun. ACM, 2011.

[4] J. Björne, F. Ginter, S. Pyysalo, J. Tsujii, T. Salakoski. Complex event extraction at PubMed scale. Bioinformatics [ISMB], 2010.

[5] P. Bohannon, N. Dalvi, Y. Filmus, N. Jacoby, S. Keerthi, A. Kirpal. Automatic web-scale information extraction. SIGMOD, 2012.

[6] M. Bundschus, M. Dejori, M. Stetter, V. Tresp, H.P. Kriegel. Extraction of semantic biomedical relations from text using conditional random fields. BMC Bioinformatics, 2008.

[7] R.J.W. Cline, K.M. Haynes. Consumer health information seeking on the Internet: the state of the art. Health education research, 2001.

[8] J. Coates. Epistemic Modality and Spoken Discourse. Transactions of the Philological Society, 1987.

[9] Z. Despotovic. Trust and Reputation in Peer-to-Peer Systems. Encyclopedia of Database Systems, 2009.

[10] X. Dong, L. Berti-Equille, Y. Hu, D. Srivastava. SOLOMON: Seeking the Truth Via Copying Detection. PVLDB, 2010.

[11] P. Ernst, C. Meng, A. Siu, G. Weikum. KnowLife: a Knowledge Graph for Health and Life Sciences. ICDE, 2014.

[12] R.E. Fan, K.W. Chang, C.J. Hsieh, X.R. Wang, C.J. Lin. LIBLINEAR: A library for large linear classification. JMLR, 2008.

[13] S. Fox, M. Duggan. Health online 2013. Pew Internet and American Life Project, 2013.
[14] S. Greene, P. Resnik. More than Words: Syntactic Packaging and Implicit Sentiment. HLT-NAACL, 2009.
[15] R.V. Guha, R. Kumar, P. Raghavan, A. Tomkins. Propagation of trust and distrust. WWW, 2004.
[16] C. Hang, Z. Zhang, M.P. Singh. Shin: Generalized Trust Propagation with Limited Evidence. IEEE Computer, 2013.
[17] IMS Institute for Healthcare Informatics. Engaging Patients through Social Media. Report, 2014, http://www.theimsinstitute.org/.
[18] K. Järvelin. J. Kekäläinen. Cumulated gain-based evaluation of IR techniques. ACM Trans. Inf. Syst., 2002.
[19] P. Jindal, D. Roth. End-to-End Coreference Resolution for Clinical Narratives. IJCAI, 2013.
[20] S.D. Kamvar, M.T. Schlosser, H. Garcia-Molina. The Eigentrust algorithm for reputation management in P2P networks. WWW, 2003.
[21] D. Koller, N. Friedman. Probabilistic Graphical Models: Principles and Techniques. MIT Press, 2009.
[22] M. Krallinger, A. Valencia, L. Hirschman. Linking Genes to Literature: Text Mining, Information Extraction, and Retrieval Applications for Biology. Genome Biology, 2008.
[23] R. Krishnamurthy, Y. Li, S. Raghavan, F. Reiss, S. Vaithyanathan. Web Information Extraction. Encyclopedia of Database Systems, 2009.
[24] X. Li, X.L. Dong, K. Lyons, W. Meng, D. Srivastava. Truth Finding on the Deep Web: Is the Problem Solved? PVLDB, 2012.
[25] X. Li, W. Meng, C.T. Yu. T-verifier: Verifying truthfulness of fact statements. ICDE, 2011.
[26] C. Lin, Y. He, R. Everson. Sentence Subjectivity Detection with Weakly-Supervised Learning. IJCNLP, 2011.
[27] C. Lin, R.C. Weng, S.S Keerthi. Trust Region Newton Method for Logistic Regression. JMLR, 2008.
[28] B. Liu. Sentiment Analysis and Opinion Mining. Synthesis Lectures on Human Language Technologies, Morgan & Claypool Publishers, 2012.
[29] A. McCallum, K. Bellare, F. Pereira. A conditional random field for discriminatively-trained finite-state string edit distance. UAI, 2005.
[30] S. Mukherjee, G. Basu, S. Joshi. Joint Author Sentiment Topic Model. SDM, 2014.
[31] S. Mukherjee, P. Bhattacharyya. Sentiment Analysis in Twitter with Lightweight Discourse Analysis. COLING, 2012.
[32] B. Pang, L. Lee. Opinion Mining and Sentiment Analysis. Foundations and Trends in Information Retrieval, 2007.
[33] J. Pasternack, D. Roth. Knowing What to Believe (when you already know something). COLING, 2010.
[34] J. Pasternack, D. Roth. Latent credibility analysis. WWW, 2013.
[35] J. Pasternack, D. Roth. Making Better Informed Trust Decisions with Generalized Fact-Finding. IJCAI, 2011.
[36] M.J. Paul, M. Dredze. Drug Extraction from the Web: Summarizing Drug Experiences with Multi-Dimensional Topic Models. HLT-NAACL, 2013.
[37] G. Peterson, P Aslani, K.A. Williams. How do consumers search for and appraise information on medicines on the Internet? A qualitative study using focus groups. Journal of Medical Internet Research, 2003.
[38] M. Recasens, C. Danescu-Niculescu-Mizil, D. Jurafsky. Linguistic Models for Analyzing and Detecting Biased Language. ACL, 2013.
[39] S. Sarawagi. Information Extraction. Foundations and Trends in Databases, 2008.
[40] C. Strapparava, A. Valitutti. Wordnet-affect: an affective extension of Wordnet. LREC, 2004.
[41] F.M. Suchanek, G. Weikum. Knowledge harvesting from text and Web sources. ICDE, 2013.
[42] C.A. Sutton, A. McCallum. An Introduction to Conditional Random Fields. Foundations and Trends in Machine Learning, 2012.
[43] P.D. Turney. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. ACL, 2002.
[44] V.G.V. Vydiswaran, C. Zhai, D. Roth. Content-driven Trust Propagation Framework. KDD, 2011.
[45] V.G.V. Vydiswaran, C. Zhai, D. Roth. Gauging the Internet Doctor: Ranking Medical Claims based on Community Knowledge. KDD Workshop on Data Mining for Healthcare, 2011.
[46] P. Westney. How to Be More-or-Less Certain in English - Scalarity in Epistemic Modality. IRAL, 1986.
[47] R.W. White, R. Harpaz, N.H. Shah, W. DuMouchel, E. Horvitz. Toward Enhanced Pharmacovigilance using Patient-Generated Data on the Internet. Nature CPT, 2014.
[48] R.W. White, E. Horvitz. From health search to healthcare: explorations of intention and utilization via query logs and user surveys. JAMIA, 2014.
[49] J. Wiebe, E. Riloff. Creating Subjective and Objective Sentence Classifiers from Unannotated Texts. CICLing, 2005.
[50] J. Wiebe, E. Riloff. Finding Mutual Benefit between Subjectivity Analysis and Information Extraction. Trans. Affective Computing, 2011.
[51] F. Wolf, E. Gibson, T. Desmet. Discourse coherence and pronoun resolution. Language and Cognitive Processes, 2004.
[52] Y. Xu, K. Hong, J. Tsujii, E.C. Chang. Feature engineering combined with machine learning and rule-based methods for structured information extraction from narrative clinical discharge summaries. JAMIA, 2012.
[53] X. Yin, J. Han, P.S. Yu. Truth discovery with multiple conflicting information providers on the Web. KDD, 2007.
[54] X. Zhu, Z. Ghahramani, J. Lafferty. Semi-supervised learning using gaussian fields and harmonic functions. ICML, 2003.