Concept Annotation from Users Perspective: A New Challenge

Souvika Sarkar
Auburn University
Alabama, USA
szs0239@auburn.edu

Shubhra Kanti (Santu) Karmaker
Auburn University
Alabama, USA
sks0086@auburn.edu

ABSTRACT

Text data is highly unstructured and can often be viewed as a complex representation of different concepts, entities, events, sentiments etc. For a wide variety of computational tasks, it is thus very important to annotate text data with the associated concepts / entities, which can put some initial structure / index on raw text data. However, It is not feasible to manually annotate a large amount of text, raising the need for automatic text annotation.

In this paper, we focus on concept annotation in text data from the perspective of real world users. Concept annotation is not a trivial task and its utility often highly relies on the preference of the user. Despite significant progress in natural language processing research, we still lack a general purpose concept annotation tool which can effectively serve users from a wide range of application domains. Thus, further investigation is needed from a user-centric point of view to design an automated concept annotation tool that will ensure maximum utility to its users. To achieve this goal, we created a benchmark corpus of two real world data-sets, i.e., "News Concept Data-set" and "Medical Concept Data-set", to introduce the notion of user-oriented concept annotation and provide a way to evaluate this task. The term "user-centric" means that the desired concepts are defined as well as characterized by the users themselves. Throughout the paper, we describe the details about how we created the data-sets, what are the unique characteristics of each data-set, how these data-sets reflect real users perspective for the concept annotation task, and finally, how they can serve as a great resource for future research on user-centric concept annotation.

CCS CONCEPTS

• **Information systems → Users and interactive retrieval**; • **Computing methodologies → Information extraction**.

KEYWORDS

Data Mining, Text annotation, Concept annotation, Data-set collection, Information retrieval, Human computer collaboration

ACM Reference Format:
Souvika Sarkar and Shubhra Kanti (Santu) Karmaker. 2022. Concept Annotation from Users Perspective: A New Challenge. In *Companion Proceedings of the Web Conference 2022 (WWW '22 Companion), April 25–29, 2022, Virtual Event, Lyon, France.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3487553.3524933

1 INTRODUCTION

As humans, we report our daily experiences as well as real world observations mostly through natural language text. Text data not only contains important details about different events around the world, but also captures subjective opinions / interpretations of the reporter about those events, which makes text data a highly valuable resource for data mining purposes. Unfortunately, text data is highly unstructured, and as a consequence, it is very difficult for computers to comprehend and process natural language efficiently. Indeed, text articles can often be viewed as a complex representation of different concepts, entities, events, sentiments, etc. For a wide variety of computational tasks, it is thus very important to annotate text data with the associated concepts / entities, which can put some initial structure / index on raw text data. However, It is not feasible to manually annotate a large amount of text, raising the need for automatic concept annotation.

Concept Annotation can be viewed as adding topic-related metadata to a text article. The idea of concept annotation is not new and several researchers have studied this problem from different perspectives in the past [28, 32, 38, 42]. Still, concept annotation is not a trivial task and its utility often relies highly on the preference of the user. Indeed, the ultimate goal of any intelligent tool is to serve the need of the end users and thus, its design principles should primarily focus on the real-world application scenarios involving the end users. Despite significant progress in natural language processing research, we still lack a general purpose concept annotation tool which can effectively serve users from a wide range of application domains. The main reason behind this limitation is the absence of a user-centric study of the concept annotation task encompassing a more realistic scenario. Thus, further investigation is needed from a user-centric point of view to design an automated concept annotation tool that ensures maximum utility to its users.

For a better demonstration of user-centric concept annotation framework, we present an intuitive example in Figure 1, where a business domain expert (the end user) is actively involved in the annotation process. Consider the domain expert (i.e. business analyst) is analyzing a large volume of financial articles and wants to computationally annotate the articles with business related concepts like "enterprise merger", "market crash", "reorganization", etc. For this real-life use case, the domain expert will provide the collection of documents (to be annotated) as well as a set of concepts which (s)he wants to be used as tags for annotating the documents. Additionally, the domain expert may also provide a list of relevant keywords

WWW '22 Companion, April 25–29, 2022, Virtual Event, Lyon, France
© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-9130-6/22/04…$15.00
https://doi.org/10.1145/3487553.3524933

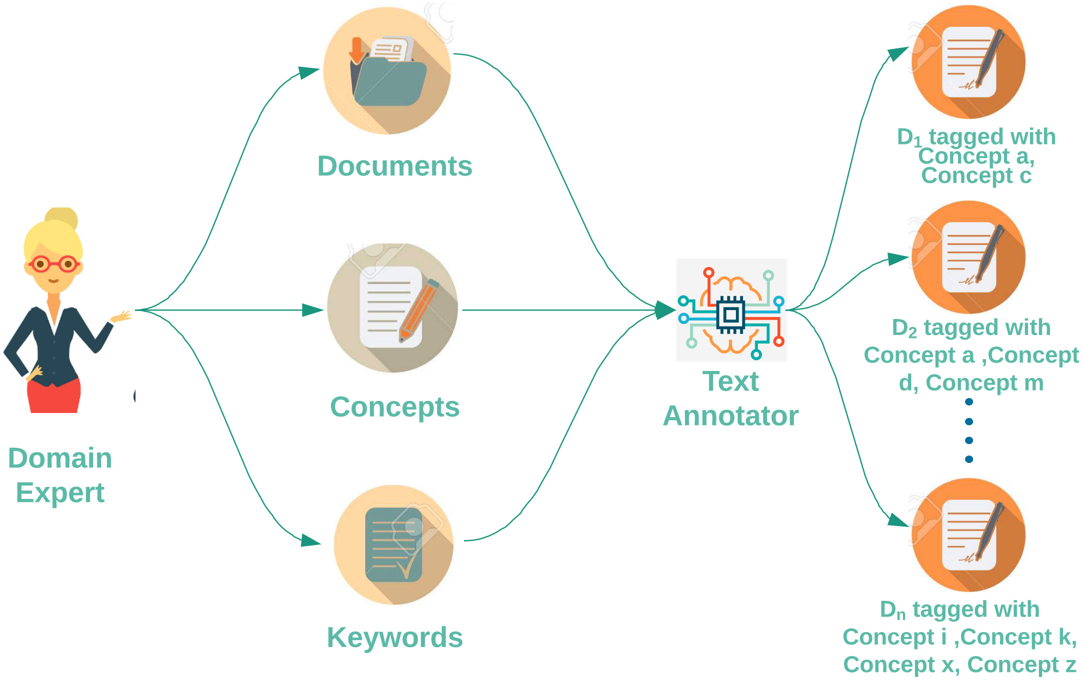

Documents
Concepts
Keywords
Domain Expert
Text Annotator
D1 tagged with Concept a, Concept c
D2 tagged with Concept a ,Concept d, Concept m
Dn tagged with Concept i ,Concept k, Concept x, Concept z

Figure 1: The user-centric concept annotation framework. The end-user (Domain Expert) provides set of documents, concepts and concept related keywords. The annotation algorithm then uses a semi-supervised approach to assign concepts to each document

associated with each concept which can be used as expert guidance for the annotation process. The automated annotation algorithm then labels each document by associating it with relevant concepts. The most important distinction of user-centric concept annotation framework from a regular one is that the desired concepts as well as concept related keywords now come from the user, who has the best knowledge of the application scenario; this provides the end user with the control to maximize the utility of the outcome of the annotation process.

In this paper, we introduce and formalize the notion of user-centric concept annotation task by introducing a benchmark corpus of two real world data-sets, i.e., “News Concept Data-set” and “Medical Concept Data-set”[1], and provide gold-standard concept labels as a way to evaluate the annotation task. The term “user-centric” means that the desired concepts are defined as well as characterized by the users themselves. The data-sets described in this paper encompasses approximately 3k articles from medical and 9k articles from the news forum.

Concept annotation is trivial when concept names are explicitly specified in the text. On the other hand, concept names do not appear directly on a frequent basis, rather they are implied throughout the text in an implicit way. Recognizing implicit concepts is an arduous job. Probing our data-sets (crawled from medical and news forum), we ascertained significant portions of the data contains these implicit concepts, hence their accurate identification, while very challenging, is critical to assure high utility for the end-users.

We conducted a preliminary study on how the domain expertise of the end user can be leveraged to mitigate the issue of implicit features. We realized that a domain expert can often provide a set of concept-related keywords and phrases from their experience. For example, a doctor can suggest words like "Stroke", "Cardiovascular", "Hypertension" etc. that are informative words for a concept like "Heart Health". In our experiments, we simulated the role of an end user (domain expert) by pre-selecting a set of concept related keywords.

In summary, we introduce and formalize the notion of user-centric concept annotation task in this paper and contribute two real world data-sets with gold annotations for evaluating this new task. Throughout the paper, we describe the details about how we created the data-sets, what are the unique characteristics of each data-set, how these data-sets reflect real users perspective for the concept annotation task, and finally, how they can serve as a great resource for future research on user-centric concept annotation.

2 RELATED WORK

Concept annotation is a fundamental research problem in NLP / text mining area and has been studied heavily in the past. As expected, many large-scale annotated corpora were created using conventional annotation schema (by a team consisting of guideline designers, annotators, and technical support staff), including Prague Dependency Treebank [6], the Arabic, English, Chinese Penn Treebank [26, 28, 42]. Another technique used for collecting high-quality annotations is to organize data challenges for the research community; for example, in the 2009 i2b2 medication challenge for concept extraction, assertion classification, and relation

[1] The resources are available at : https://1drv.ms/f/s!Aiv6VuLp2LFnaTvBRSPAZ4xRJkg

classification task, [40] was created by the i2b2 organizers and the participating teams.

In the biomedical domain, semantically annotated corpora including GENIA corpus [23] and PennBioIE corpus [27] are publicly available. Kulick et al. [24] presented annotation guidelines for Biomedical information extraction. Xia et al. [43] presented three corpora for clinical NLP studies. One of them identifies critical recommendations in radiology reports, and the other two indicate whether a patient has pneumonia based on chest X-ray reports or ICU reports. Cohen et al. [11] discussed about design features, characteristics and contributed in design ideas for biomedical corpora. Bijoy et.al. [5] performed simple keyword-based annotation on COVID-19 related tweets to analyze frequent symptoms associated with the pandemic.

Another school of researchers focused on concept annotation tasks for the legal domain. For example, Dragoni et al. [14] and Wyner et al. [41] worked on rule extraction from legal documents using Natural Language Processing techniques. Soriaet al. [36] and Spinosa et al. [37] performed semantic analysis of the textual amendments and extracted metadata / regulatory content. Biagioli et.al. [4] studied retrieval of norms from legal documents using NLP methods.

Besides introducing new data-sets for concept annotation task, researchers have developed several automated tools to perform the annotation itself. For example, BRAT [38] is a web-based Tool for assisted text annotation featuring high-quality annotation visualization, intuitive annotation interface and support. Knowtator [32] is a general-purpose text annotation tool which can aid the manual creation of annotated corpora that can be used for evaluating or training a variety of natural language processing systems. Kalina et al. developed GATE Teamware [7] which is an open-source, web-based, collaborative text annotation framework. It facilitates to carry out complex corpus annotation projects, involving distributed annotator teams. The tool comes with different user roles such as annotator, manager, to support the complex workflows and user interactions that usually occur in corpus annotation projects. Seeker [13] is a platform for large-scale text analytics, and SemTag is an application written on the platform to perform automated semantic tagging of large corpora.

One closely similar task to concept annotation is the Named entity recognition (NER) task, where the goal is to identify references to real-world entities mentioned in raw text data. Kulkarni et al. [25] in their paper presented annotation of Wikipedia Entities in Web Text. Their method annotated Web pages with entities from an entity catalog, such as Wikipedia. Another school of researchers aimed to automatically cross-reference significant terms with Wikipedia [16, 31]. They used NLP techniques to annotate terms within the text that are short, improperly formed, and also unstructured, and enhanced it with links to the appropriate Wikipedia articles. Mihalcea et al. [29] showed that given an input document, their system can identify the important concepts in the text and link these concepts to the corresponding Wikipedia pages.

A probabilistic view, as provided by topic models, performs modestly for identifying concepts in unstructured data. Multiple research [8, 39] has shown it is possible to learn to annotate from well-annotated collections of metadata through supervised learning. Iwata et.al. [20] proposed a topic model for analyzing and excerpting content related annotations from noisy annotated discrete data such as web pages stored in bookmarks. Poursabzi-sangdeh et al. [34] merged document classification and topic models, where topic modeling was used to uncover the underlying semantic structure of documents in the collection. Engels et al. [15] proposed an automatic annotation scheme, in which they employed a latent topic model to generate topic distributions given a video and associated text. Karmaker et al. [22] proposed a generative feature-topic model that can mine implicit features from online reviews, through unsupervised statistical learning.

Sentiment analysis is another closely related area which often benefit from the concept annotation task as sentiments are often expressed at the concept / event level. Erik Cambria et al. [10] mentioned that sentiment analysis is a suitcase research problem that requires undertaking several NLP tasks, in particular 15 tasks including "Concept Annotation". In [9], authors extracted topics / concepts that are highly correlated with the positive and negative sentiments (from opinions). In [2, 3], researchers have presented strategy for automatic sentiment analysis and concept labeling over Spanish Twitter data. Hassan et.al. [18] have developed an automatic sexual violence report tracking system by extensively annotating tweets with #metoo hashtag.

In contrast to all studies discussed above, our strategy takes a different perspective to this classic problem, i.e, focusing on real-world use-case scenarios. Ad-hoc user requirements / preferences for concept annotation can be supported in our problem formulation by actively engaging the user in the process and allowing them to provide their own desired set of concepts and keywords.

3 PROBLEM STATEMENT

The goal task is to annotate a collection of documents D with a set of concepts C, where each concept C has a list of associated keywords K_C provided by the domain expert / end user. Our user-centric problem set-up assumes that the end user provides all the documents, concepts and keyword-lists as inputs. The user here is usually a domain expert with specialized knowledge or skills in a particular area of endeavor (e.g., a cardiologist is an expert in the domain of “heart health”).

Let D = $\{d_1, d_2, ..., d_n\}$ be the collection of documents where each d_i represents a document in the corpus. Let C = $\{c_1, c_2, ..., c_m\}$ is the collection of all concepts the user is interested in. Each concept is represented by a word/phrase and is associated with a set of related keywords $K_c = \{k_c^1, k_c^2, ..., k_c^p\}$. The goal task is to annotate each $d_i \in D$ with a set of concepts, $\widetilde{C} \subseteq C$. Noteworthy, a document may have multiple concepts associated with it as well as a concept may be associated with multiple documents. Moreover, a concept can be roughly characterized by a set of keywords and different sets of keywords may characterize different concepts.

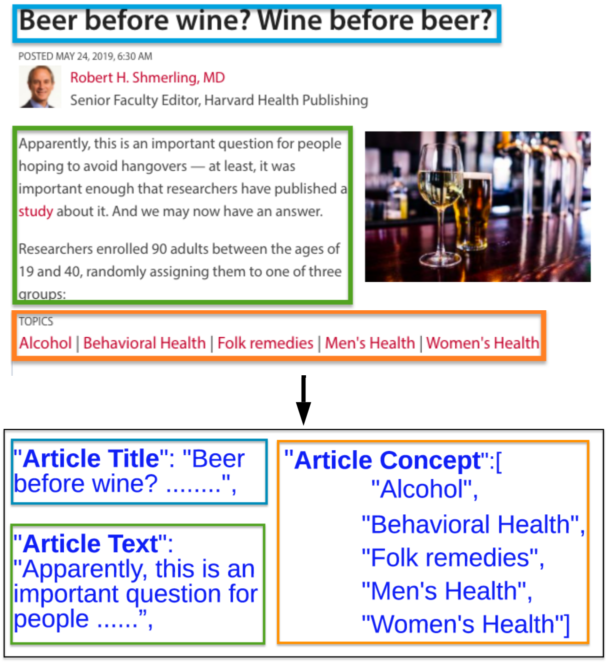
A concept c_k may not occur by it's name / phrase explicitly in a document d_i. For example, a document about “Mental Health” may not include the exact phrase “Mental Health”, but still talk about "Depression", "Anxiety" and "Antidepressant Drugs". Thus, the concept “Mental Health” is implicit in this document and it is equally important to annotate the implicit concepts within a document as well as the explicit concepts. To help us tackle this problem, a domain expert can create a keyword dictionary for each concept to reduce the number of implicit mentions and convert them into explicit mentions. For example, a doctor can provide the list {"Bone," "Calcium," "Fractures"} as keywords for the concept "Osteoporosis" and if we find that there are explicit mentions of these words in a document, it can be considered as revealing the concept "Osteoporosis" explicitly. Below, we formalize the input and output of our user-centric concept annotation task.

Input: a collection of documents D = $\{d_1, d_2, ..., d_n\}$, a collection of concepts C = $\{c_1, c_2, ..., c_m\}$ and a set of keywords K_c = $\{k_c^1, k_c^2, ..., k_c^p\}$ associated with each concept c.

Output: $\left\{\gamma_{d_i}^{c_j}\right\}$ for each $d_i \in D$ and $c_j \in C$, where, $\gamma_{d_i}^{c_j} = 1$ if concept c_j is present in document d_i, or 0 otherwise.

4 THE NEW RESOURCE

This section describes the two new data-sets we created for user-centric concept annotation, i.e., "News Concept" and "Medical Concept" Data-sets. Below, we discuss how we created these data-sets step-by-step and highlight some challenges we faced along the way.



Beer before wine? Wine before beer?

POSTED MAY 24, 2019, 6:30 AM

Robert H. Shmerling, MD

Senior Faculty Editor, Harvard Health Publishing

Apparently, this is an important question for people hoping to avoid hangovers — at least, it was important enough that researchers have published a study about it. And we may now have an answer.

Researchers enrolled 90 adults between the ages of 19 and 40, randomly assigning them to one of three groups:

TOPICS

Alcohol | Behavioral Health | Folk remedies | Men's Health | Women's Health

"Article Title": "Beer before wine? ....",

"Article Text": "Apparently, this is an important question for people ......",

"Article Concept":["Alcohol", "Behavioral Health", "Folk remedies", "Men's Health", "Women's Health"]

Figure 2: Sample document from Medical Data-set: Article scraped from www.health.harvard.edu in JSON file in the form of '*Article Title*', '*Article Text*', '*Article Concept*'

4.1 Data Collection and cleaning

A collection of publicly available online news and medical-blog articles were crawled from the web to create our Data-sets. Each article was already tagged with one or more concepts by human annotators. For example an article titled, "Why eating slowly may help you feel full faster" is associated with concepts "Diet and Weight Loss" and "Health". We scraped the article titles, article texts, and article concepts from the news and medical-blog websites and and stored them as JSON objects. Figure 2 shows how each article is stored in the JSON file with keys: "article title", "article text" and "article concept". Blue box highlights article title, Green box highlights article text, and Orange box highlights topics which we considered as concepts.

Data-set ->	News Concept	Medical Concept
URL	newsbusters.org	health.harvard.edu
Total # of Articles	8940	2066
# of Original Concepts	7199	2331
# of Concepts Retained	12	18
Avg. # of concepts per article with ≥ 1 concept	1.29	1.47

Table 1: An overview of the new Data-sets

Domain	Concept	Merged Concepts
Medical	Arthritis	Arthritis, Osteoarthritis
Medical	Children's Health	Children's Health, Parenting
Medical	Headache	Headache, Migraines
Medical	Healthy Eating	Healthy Eating, Diet and Weight Loss
Medical	Heart Health	Heart Health, Hypertension and Stroke
Medical	Mental Health	Mental Health, Anxiety and Depression, Stress
Medical	Prostate Knowledge	Prostate Knowledge, Prostate Health, Living With Prostate Cancer
Medical	Women's Health	Women's Health, Family Planning and Pregnancy, Pregnancy
News	2020 Presidential	2020 Presidential, Campaigns and Elections
News	Celebrities	Celebrities, Hollywood, Movies
News	Economy	Economy, Recession, Budget, Stock Market, Banking/Finance, Capitalism
News	Religion	Religion, Christianity, Anti-Religious Bias
News	Sexuality	Sexuality, Homosexuality, Sexism, Same sex marriage, Transgender
News	Trump-Russia probe	Trump-Russia probe, Mueller Report

Table 2: Details of merged concepts for Medical and News data-set

As part of data cleaning, we observed few overlapping concepts that were mostly appearing together, so we merged those similar/ overlapping concepts into a single concept. Through meticulous manual effort, we then selected a subset of the total available (merged) concepts based on the following course of actions: 1) Removing duplicate concepts, 2) Ignoring entities like people, place etc. to be considered as concepts, 3) Discarding very general concepts like sports, politics etc, 4) Removing concepts with a low frequency of associated articles and 5) Selecting a subset of concepts that can ensure a high level of diversity within each domain. The final corpora contains 12 unique concepts for News Data-set and 18 unique concepts for Medical Data-set, an overview of which are presented in table 1. The statistics of the merged concepts for both data-sets are presented in table 2.

4.2 Prevalence of Implicit Mentions

For each article, we checked whether the ground-truth concepts can be identified by performing a simple Boolean check with the concept name. Since our data-sets are comprised of lengthy articles and each article is a complex representation of various concepts, entities and events, only checking the concept name in the text performed poorly. We report the results of this simple Boolean matching based annotation technique in tables 3 and 4 for News and Medical Data-sets, respectively. For each article, the concepts assigned by the simple Boolean approach were compared against the human annotated concept to compute the true positive, false positive and false negative statistics, which are defined as below:

- **True Positive:** Number of concepts correctly extracted.
- **False Negative:** Number of concepts not extracted.
- **False Positive:** Number of concepts incorrectly extracted.

Concept Name	Total Count	True Positive	False Negative	False Positive
2020 Presidential	2212	147	2065	57
Abortion	411	385	26	337
Celebrities	497	46	451	107
Coronavirus	227	221	6	84
Economy	317	148	169	323
Foreign Policy	370	41	329	126
Global Warming	312	128	184	33
Immigration	374	249	125	313
Religion	327	89	238	116
Sexuality	689	90	599	40
Trump Impeachment	1085	34	1051	15
Trump-Russia probe	466	17	449	3

Table 3: News Forum: Concept annotation result based on Boolean check by the concept name

Concept Name	Total Count	True Positive	False Negative	False Positive
Addiction	95	64	31	36
Alcohol	9	9	0	237
Arthritis	46	40	6	96
Brain and cognitive health	92	0	92	0
Breast Cancer	27	24	3	34
Cancer	172	164	8	389
Children's Health	290	0	290	0
Exercise and Fitness	176	0	176	0
Headache	31	29	2	120
Healthy Eating	313	44	269	11
Heart Health	255	33	222	20
Mental Health	300	103	197	72
Osteoporosis	23	19	4	39
Pain Management	98	10	88	4
Prostate Knowledge	161	4	157	0
Sleep	58	58	0	270
Smoking cessation	15	4	11	8
Women's Health	172	0	172	0

Table 4: Medical Forum: Concept annotation result based on Boolean check by the concept name

Extrapolating the very high 'False Negative' values in Table 3 and 4, we concluded that many of the concepts are not explicitly mentioned in the article and are thus "Implicit concepts". The difference between the two can be further clarified through an example. We consider a concept as explicit if the concept names are explicitly mentioned in the article text. For example, the following sentence is from an article related to concept **Corona virus**, *"Americans should feel much better about the corona virus coming under control"*, which mentions the concept **Corona virus** explicitly in the text body. Whereas, for implicit concepts, the concept name is not directly mentioned in the article text, rather the concept is somewhat implied. For example, the following sentence is taken from an article annotated with the concept **Women's Health**, *"Studies question ban on alcohol during pregnancy."* Here, the text does not contain the phrase **Women's Health**, yet a human can easily relate it to the same concept. We consider these cases as implicit mentions on the target concept. Based on the above observation, we performed a detailed analysis regarding explicit and implicit concepts for both data-sets, the results of which are presented in table 5. Due to the abundance of these implicit concepts as reported in Table 5, we conclude that simply checking the concept name in the text will not yield a high quality automatic annotation.

5 BRINGING THE USER INTO THE LOOP

To mitigate the issue of the ubiquity of implicit concepts, we started delving into the data-sets for finding alternative approaches. On further assessment, we realized that in cases where concept names are not directly mentioned in the text, some informative keywords related to the concept are always present in the article text. Indeed, each concept can be conceptually viewed as a cloud of its informative keywords and different concepts will essentially yield different word clouds. More interestingly, these informative keywords (word cloud) can be provided by the end user (domain expert) conducting the annotation task. In fact, we realized this is what mostly happens in real-world cases and decided to simulate this case artificially while creating our data-sets. The whole simulation process can be summarised in the following 2 steps.

Step 1- Extracting Informative keywords: We extracted the informative keywords for each concept using the TF-IDF (Term Frequency - Inverse Document Frequency) heuristics on the documents tagged with each concepts [See Algorithm 1 for details]. For example, the articles related to concept 'Heart Health' yielded informative keywords like 'Cardiovascular', 'Stroke', 'Heart attack', 'Blood pressure', 'Cholesterol', 'Heart' etc (refer to figure 3). This way, we prepared a JSON file with the list of concepts and respective informative keywords.

Statistic	News Forum	Medical Forum
Total Explicit mentions	1577	604
Total Implicit mentions	5622	1727
Percentage of Explicit mentions	21.91	25.92
Percentage of Implicit mentions	78.09	74.08
Avg Explicit mentions per article with ≥ 1 concept	0.28	0.38
Avg Implicit mentions per article with ≥ 1 concept	1.01	1.09
Number of articles without single concept	3401	482

Table 5: Details of articles from News and Medical Forum

Step 2 - Building a Keyword Dictionary: For each concept, we selected three (empirically set) informative keywords from the keyword list extracted during Step 1. This selection was done through careful manual inspection in order to simulate a real-life end user. For example, for concept "Global Warming", the following keywords were chosen: {'Climate', 'Planet', 'Green'}, which are intuitively related to the concept. Similarly, keywords for Religion includes words like {'Church','Christian', 'Religious'}. Tables 6 and 7 contain concepts and corresponding keywords details from medical and news data-set respectively.

Algorithm 1 Pseudocode for informative keyword extraction

```
Input: A concept for which keyword will be extracted, set of article text and corresponding concept.
Output: list of keywords
for each article concept do
    if input concept present in article associated concept then add article text to a list
    end if
end for
Call TF-IDF function
        Pass In:Extracted articles list
        Pass Out:Keywords
end
```

Figure 3: Set of keywords relevant to the concept Global Warming and Heart Health

6 A RUDIMENTARY KEYWORD-BASED ANNOTATION ALGORITHM

To discover concepts by using the informative keyword list provided by the user, we experimented with a rudimentary keyword-Based annotation algorithm (See Algorithm 2). To measure the performance of this rudimentary approach, we use popular measures available in the literature: Precision, Recall, F1 measure and False Positive Rate. The corresponding True Positive, False Positive and, false negative values were calculated based on pseudo code presented in algorithm 3.

Algorithm 2 Pseudo code for concept annotation

```
Input: Article text, Article title and JSON file containing concept names,keywords
Output: Articles tagged with concepts ;
for each article text do
    check whether concept name or any one of the informative keywords are present or not in corresponding article text
    if present then label the article with the concept
    end if
end for
```

Results of this rudimentary keyword-based annotation algorithm is shown in table 8 and 9 for News and Medical Data-set Respectively. For instance, for the concept 'Healthy Eating' in our medical data-set, keyword based annotation approach obtained a precision of 0.37, recall of 0.94, corresponding F1 measure of 0.53 and a False Positive Rate of 0.29. The distribution of F1 Measure and FP rate for each concept has been shown in figure 4 and 5.

Concept Name	Keywords
Addiction	Opioids, Alcohol, Drug
Alcohol	Wine, Consumption, Sud
Arthritis	Pain, Knee, Joint
Brain and cognitive health	Brain, Dementia, Memory
Breast Cancer	Mastectomy, Mammograms, Prophylactic
Cancer	Screening, Radiation, Cells
Children's Health	Parents, Children, Babies
Exercise and Fitness	Exercise, Activity, Physical
Headache	Migraine, Sinus, Chronic pain
Healthy Eating	Diet, Foods, Weight
Heart Health	Hypertension, Stroke, Cardiovascular
Mental Health	Depression, Anxiety, Antidepressant
Osteoporosis	Bone, Calcium, Fractures
Pain Management	Opioid, Pain, Osteoarthritis
Prostate Knowledge	Prostate, Psa, Screening
Sleep	Night, Apnea, Insomnia
Smoking cessation	Cigarettes, Smoking, Vaping
Women's Health	Pregnancy, Breast, Birth

Table 6: Concepts and related keywords from Medical forum

Concept Name	Keywords
2020 Presidential	Trump, Biden, Campaign
Abortion	Parenthood, Baby, Court
Celebrities	Hollywood, Actor, Movies
Coronavirus	Virus, Covid, Covid 19
Economy	Recession, Budget, Stock Market
Foreign Policy	Iran, Soleimani, Security
Global Warming	Climate, Planet, Green
Immigration	Border, Immigrants, Detention
Religion	Christian, Religious, Church
Sexuality	Gay, Lgbtq, Transgender
Trump Impeachment	Trump, Impeachment, Democrats
Trump-Russia probe	Mueller, Russia, Trump

Table 7: Concepts and related keywords from News forum

Apparently, simple keywords search appeared to achieve better results than using the Boolean concept name matching technique and yielded higher True Positive and lower False Negative numbers. However, we also observed false positive counts on the higher end, meaning this approach may not be very useful in practical applications where precision is a high priority. Further, the performance of the rudimentary approach greatly depends on the choice of keywords; without appropriate keywords, the approach may suffer seriously.

7 FUTURE USAGE OF THE RESOURCE

In this paper, we introduced the notion of user-centric concept annotation task and created 2 data-sets for this new challenge, so that researchers can dig deeper into this important problem. we discussed how we created these data-sets step-by-step including data collection, cleaning, implicit feature identification and end-user

Algorithm 3 Pseudo code for finding True Positive, False Positive, False Negative count

```
Input: Annotated dataset, a concept name;
Output: Total, True Positive, False Positive, False Positive count for a given concept ;
for each article do
    if input concept present in 'Explicit Article Concept' field then count total number
    end if
    if input concept present in both field, 'Explicit Article Concept and 'Article Concept' then count as TRUE POSITIVE
    end if
    if input concept present in 'Explicit Article Concept' but not in 'Article Concept' then count as FALSE POSITIVE
    end if
    if input concept not present in 'Explicit Article Concept but in 'Article Concept' then count as FALSE NEGATIVE
    end if
end for
```

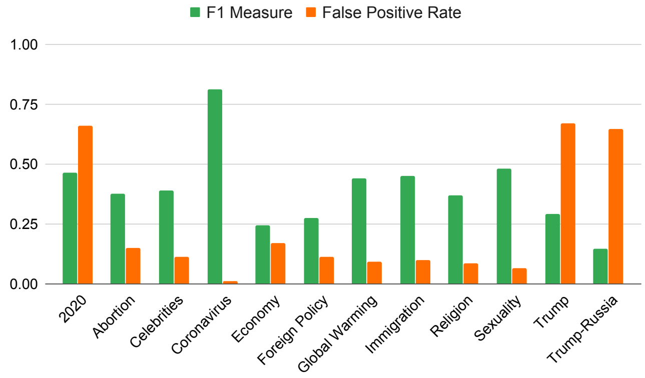


Figure 4: Performance measures of News forum concept annotation process

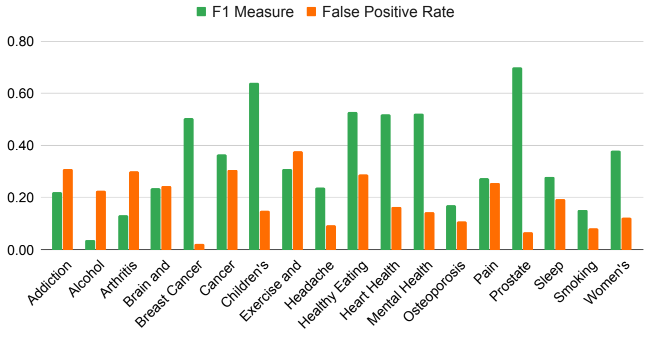


Figure 5: Performance measures of Medical forum concept annotation process

simulation. Based on the our analysis, we make some important observations about the new data-sets. First, direct searching for concept name yields very low recall, thus it is not useful. The rudimentary keyword-based annotation algorithm performs better in terms of recall, but at the expense of low precision and high false positive rate, and consequently low F1 score. Therefore, more sophisticated methods need to be devised to achieve a reasonable accuracy. To facilitate research in this direction, we have published the data-sets, scripts for data-loading and data-statistics computation and a readme file with detailed instructions on how to use this resource, upon acceptance of our submitted manuscript. Below we point towards possible research directions using this resource.

Exploring Semantic Embeddings Vectors: Word embedding techniques like Word2Vec [30] and Glove [33] can be quite handy for concept annotation tasks. Less frequent words in the text corpora which do not display strong correlations with other words may greatly benefit from such embedding representations because embeddings are pre-learnt from a big corpus of text and expected to have a more robust representation for less frequent words in the text being annotated. This ability to represent words, phrases as vectors as well as represent similar words closely in vector space, may lead Word2vec to produce very promising results if used in text annotations.

Exploring Deep Sequential Models: Text data is sequential in nature and thus, concept annotation in text can be viewed as a sequence labeling task. From this perspective, Deep Sequential Models like Recurrent Neural Network (RNN), Long Short Term Memory (LSTM [19], TILM [35]), Transformers (BERT [12]) can used for performing concept annotation. The reason being, long sequences often play a vital role for context understanding and concept identification.

Exploring Constrained Topic Modeling: Topic modeling techniques [1] (LDA, PLSA, NNMF) are popular unsupervised techniques for discovering the abstract "topics" from a collection of documents. However, topic modeling techniques are not directly applicable for user-centric concept annotation task as it is a semi-supervised task with active engagement from the user, while topic models are completely unsupervised. Thus, a user-preference based constrained topic modeling technique needs to devised for the annotation task.

Adding External Knowledge Graphs: Knowledge Graphs are popular techniques for capturing relationships among entities and concepts [17]. Thus, external knowledge graphs can help identify ambiguous concepts by exploiting the internal graph relations and mapping them on the text document being annotated, which is definitely a promising future research direction.

8 CONCLUSION

It is evident that in the era of web scale unstructured data, annotation is a crucial process. Information retrieval and Knowledge mining becomes much easier if data is categorized and annotated precisely. Consequently, a general annotation tool which can effectively serve end users from a wide area of application domain will greatly benefit the movement of data driven design and discovery. For example, an annotated medical data-set can promptly retrieve similar "Cancer" cases from the past, an annotated news data-set can be useful in retrieving all "Presidential Election" news in a few clicks. Annotated systems will aid enterprises to store and retrieve digital information efficiently, which will accelerate all kinds of data driven decision process [21].

With the rapid growth of Big-data, it is infeasible to perform manual annotation, as it is slow and expensive. The ever-increasing

Concept	Total Count	True Positive	False Negative	False Positive	Precision	Recall	F1 Measure	FP Rate
2020 Presidential	2212	2011	201	4444	0.31	0.91	0.46	0.66
Abortion	411	395	16	1288	0.23	0.96	0.38	0.15
Celebrities	497	345	152	936	0.27	0.69	0.39	0.11
Coronavirus	227	225	2	102	0.69	0.99	0.81	0.01
Economy	317	248	69	1465	0.14	0.78	0.24	0.17
Foreign Policy	370	214	156	975	0.18	0.58	0.27	0.11
Global Warming	312	309	3	785	0.28	0.99	0.44	0.09
Immigration	374	353	21	844	0.29	0.94	0.45	0.10
Religion	327	239	88	732	0.25	0.73	0.37	0.08
Sexuality	689	391	298	542	0.42	0.57	0.48	0.07
Trump Impeachment	1085	1081	4	5254	0.17	1.00	0.29	0.67
Trump-Russia probe	466	462	4	5466	0.08	0.99	0.14	0.65

Table 8: Details of concepts and number of True Positive, False Negative and False Positive articles from News forum

Concept	Total Count	True Positive	False Negative	False Positive	Precision	Recall	F1 Measure	FP Rate
Addiction	95	88	7	610	0.13	0.93	0.22	0.31
Alcohol	9	9	0	464	0.02	1.00	0.04	0.23
Arthritis	46	46	0	606	0.07	1.00	0.13	0.30
Brain and cognitive health	92	76	16	480	0.14	0.83	0.23	0.24
Breast Cancer	27	24	3	44	0.35	0.89	0.51	0.02
Cancer	172	168	4	580	0.22	0.98	0.37	0.31
Children's Health	290	263	27	267	0.50	0.91	0.64	0.15
Exercise and Fitness	176	162	14	714	0.18	0.92	0.31	0.38
Headache	31	30	1	192	0.14	0.97	0.24	0.09
Healthy Eating	313	293	20	506	0.37	0.94	0.53	0.29
Heart Health	255	194	61	300	0.39	0.76	0.52	0.17
Mental Health	300	195	105	252	0.44	0.65	0.52	0.14
Osteoporosis	23	23	0	223	0.09	1.00	0.17	0.11
Pain Management	98	95	3	502	0.16	0.97	0.27	0.26
Prostate Knowledge	161	154	7	129	0.54	0.96	0.69	0.07
Sleep	72	72	0	385	0.16	1.00	0.28	0.19
Smoking Cessation	15	15	0	168	0.08	1.00	0.15	0.08
Women's Health	172	95	77	233	0.29	0.55	0.38	0.12

Table 9: Details of concepts and number of True Positive, False Negative and False Positive articles from Medical forum

scale of the data in different areas, new types of content, creates an ever-growing need to continuously adapt and refine annotation methods. Although the area of text annotation is not in the nascent phase, it has not been well-studied from a user-centric point of view. Our contribution in this area will enable further research including novel machine learning practices and information retrieval systems. We strongly believe that an interdisciplinary effort from multiple research areas including natural language processing, Human-Computer Interaction, Machine Learning and Information Retrieval is needed to effectively design a general-purpose user-centric concept annotation tool. Therefore, we encourage the community to make use of the corpora for solving this fundamental yet crucial data science task.

9 ETHICS STATEMENT

In this paper, we have discussed the creation of two benchmark data-sets from real-world user generated contents. To the fulfilment of this goal, we have scraped contents from 2 different publicly accessible websites. Hence, we did not obtain any explicit approval as our intended use of the contents is entirely educational/research-focused and the created data-sets will only be shared with other researchers for research purposes exclusively. We have not tried to identify any private information from the collected data in any way which can result in a privacy violation. In the whole experiment, we only used open source packages and libraries, along with proper citations as required.

REFERENCES

[1] Rubayyi Alghamdi and Khalid Alfalqi. 2015. A survey of topic modeling in text mining. *Int. J. Adv. Comput. Sci. Appl.(IJACSA)* 6, 1 (2015).

[2] Antonio Fernández Anta, Luis Núnez Chiroque, Philippe Morere, and Agustín Santos. 2013. Sentiment analysis and topic detection of Spanish tweets: A comparative study of of NLP techniques. *Procesamiento del lenguaje natural* 50 (2013), 45–52.

[3] Fernando Batista and Ricardo Ribeiro. 2013. Sentiment analysis and topic classification based on binary maximum entropy classifiers. *Procesamiento del lenguaje natural* 50 (2013), 77–84.

[4] Carlo Biagioli, Enrico Francesconi, Andrea Passerini, Simonetta Montemagni, and Claudia Soria. 2005. Automatic semantics extraction in law documents. In *Proceedings of the 10th international conference on Artificial intelligence and law.* 133–140.

[5] Biddut Sarker Bijoy, Syeda Jannatus Saba, Souvika Sarkar, Md Saiful Islam, Sheikh Rabiul Islam, Mohammad Ruhul Amin, and Shubhra Kanti Karmaker Santu. 2021. COVID19α: Interactive Spatio-Temporal Visualization of COVID-19 Symptoms through Tweet Analysis. In *26th International Conference on Intelligent User Interfaces-Companion*. 28–30.

[6] Alena Böhmová, Jan Hajič, Eva Hajičová, and Barbora Hladká. 2003. The Prague dependency treebank. In *Treebanks*. Springer, 103–127.

[7] Kalina Bontcheva, Hamish Cunningham, Ian Roberts, Angus Roberts, Valentin Tablan, Niraj Aswani, and Genevieve Gorrell. 2013. GATE Teamware: a web-based, collaborative text annotation framework. *Language Resources and Evaluation* 47, 4 (2013), 1007–1029.

[8] Markus Bundschus, Volker Tresp, and Hans-Peter Kriegel. 2009. Topic models for semantically annotated document collections. In *NIPS workshop: Applications for Topic Models: Text and Beyond*. 1–4.

[9] Keke Cai, Scott Spangler, Ying Chen, and Li Zhang. 2010. Leveraging sentiment analysis for topic detection. *Web Intelligence and Agent Systems: An International Journal* 8, 3 (2010), 291–302.

[10] Erik Cambria, Soujanya Poria, Alexander Gelbukh, and Mike Thelwall. 2017. Sentiment analysis is a big suitcase. *IEEE Intelligent Systems* 32, 6 (2017), 74–80.

[11] K Bretonnel Cohen, Lynne Fox, Philip Ogren, and Lawrence Hunter. 2005. Corpus design for biomedical natural language processing. In *Proceedings of the ACL-ISMB workshop on linking biological literature, ontologies and databases: mining biological semantics*. 38–45.

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[13] Stephen Dill, Nadav Eiron, David Gibson, Daniel Gruhl, Ramanathan Guha, Anant Jhingran, Tapas Kanungo, Sridhar Rajagopalan, Andrew Tomkins, John A Tomlin, et al. 2003. SemTag and Seeker: Bootstrapping the semantic web via automated semantic annotation. In *Proceedings of the 12th international conference on World Wide Web*. 178–186.

[14] Mauro Dragoni, Serena Villata, Williams Rizzi, and Guido Governatori. 2016. Combining NLP approaches for rule extraction from legal documents.

[15] Chris Engels, Koen Deschacht, Jan Hendrik Becker, Tinne Tuytelaars, Sien Moens, and Luc J Van Gool. 2010. Automatic annotation of unique locations from video and text.. In *BMVC*. 1–11.

[16] Paolo Ferragina and Ugo Scaiella. 2010. Tagme: on-the-fly annotation of short text fragments (by wikipedia entities). In *Proceedings of the 19th ACM international conference on Information and knowledge management*. 1625–1628.

[17] Larry González and Aidan Hogan. 2018. Modelling dynamics in semantic web knowledge graphs with formal concept analysis. In *Proceedings of the 2018 World Wide Web Conference*. 1175–1184.

[18] Naeemul Hassan, Amrit Poudel, Jason Hale, Claire Hubacek, Khandaker Tasnim Huq, Shubhra Kanti Karmaker Santu, and Syed Ishtiaque Ahmed. 2020. Towards automated sexual violence report tracking. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 14. 250–259.

[19] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[20] Tomoharu Iwata, Takeshi Yamada, and Naonori Ueda. 2009. Modeling social annotation data with content relevance using a topic model. In *Advances in Neural Information Processing Systems*. 835–843.

[21] Shubhra Kanti Karmaker Santu, Chase Geigle, Duncan Ferguson, William Cope, Mary Kalantzis, Duane Searsmith, and Chengxiang Zhai. 2018. SOFSAT: Towards a Setlike Operator based Framework for Semantic Analysis of Text. *ACM SIGKDD Explorations Newsletter* 20, 2 (2018), 21–30.

[22] Shubhra Kanti Karmaker Santu, Parikshit Sondhi, and ChengXiang Zhai. 2016. Generative feature language models for mining implicit features from customer reviews. In *Proceedings of the 25th ACM international on conference on information and knowledge management*. 929–938.

[23] J-D Kim, Tomoko Ohta, Yuka Tateisi, and Jun'ichi Tsujii. 2003. GENIA corpus—a semantically annotated corpus for bio-textmining. *Bioinformatics* 19, suppl_1 (2003), i180–i182.

[24] Seth Kulick, Ann Bies, Mark Liberman, Mark Mandel, Ryan McDonald, Martha Palmer, Andrew Schein, Lyle Ungar, Scott Winters, and Peter White. 2004. Integrated annotation for biomedical information extraction. In *HLT-NAACL 2004 Workshop: Linking Biological Literature, Ontologies and Databases*. 61–68.

[25] Sayali Kulkarni, Amit Singh, Ganesh Ramakrishnan, and Soumen Chakrabarti. 2009. Collective annotation of Wikipedia entities in web text. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. 457–466.

[26] Mohamed Maamouri and Ann Bies. 2004. Developing an Arabic treebank: Methods, guidelines, procedures, and tools. In *Proceedings of the Workshop on Computational Approaches to Arabic Script-based languages*. 2–9.

[27] Mark A Mandel. 2006. Integrated annotation of biomedical text: creating the PennBioIE corpus. In *Proceedings of the Workshop on Text Mining, Ontologies and Natural Language Processing in Biomedicine, Manchester, UK, 2006*.

[28] Mitchell Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. (1993).

[29] Rada Mihalcea and Andras Csomai. 2007. Wikify! Linking documents to encyclopedic knowledge. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*. 233–242.

[30] 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*. 3111–3119.

[31] David Milne and Ian H Witten. 2008. Learning to link with wikipedia. In *Proceedings of the 17th ACM conference on Information and knowledge management*. 509–518.

[32] Philip Ogren. 2006. Knowtator: a protégé plug-in for annotated corpus construction. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Demonstrations*. 273–275.

[33] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

[34] Forough Poursabzi-Sangdeh and Jordan Boyd-Graber. 2015. Speeding Document Annotation with Topic Models. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*. 126–132.

[35] Shubhra Kanti Karmaker Santu, Kalyan Veeramachaneni, and Chengxiang Zhai. 2019. TILM: Neural language models with evolving topical influence. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*. 778–788.

[36] Claudia Soria, Roberto Bartolini, Alessandro Lenci, Simonetta Montemagni, and Vito Pirrelli. 2007. Automatic extraction of semantics in law documents. In *Proceedings of the V Legislative XML Workshop*. European Press Academic Publishing, 253–266.

[37] PierLuigi Spinosa, Gerardo Giardiello, Manola Cherubini, Simone Marchi, Giulia Venturi, and Simonetta Montemagni. 2009. NLP-based metadata extraction for legal text consolidation. In *Proceedings of the 12th International Conference on Artificial Intelligence and Law*. 40–49.

[38] Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. BRAT: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 102–107.

[39] Suppawong Tuarob, Line C Pouchard, Prasenjit Mitra, and C Lee Giles. 2015. A generalized topic modeling approach for automatic document annotation. *International Journal on Digital Libraries* 16, 2 (2015), 111–128.

[40] Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2011. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association* 18, 5 (2011), 552–556.

[41] Adam Z Wyner and Wim Peters. 2011. On Rule Extraction from Regulations.. In *JURIX*, Vol. 11. 113–122.

[42] Fei Xia, Martha Palmer, Nianwen Xue, Mary Ellen Okurowski, John Kovarik, Fu-Dong Chiou, Shizhe Huang, Tony Kroch, and Mitchell P Marcus. 2000. Developing Guidelines and Ensuring Consistency for Chinese Text Annotation.. In *LREC*.

[43] Fei Xia and Meliha Yetisgen-Yildiz. 2012. Clinical corpus annotation: challenges and strategies. In *Proceedings of the Third Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM'2012) in conjunction with the International Conference on Language Resources and Evaluation (LREC), Istanbul, Turkey*.