A Machine Learning Approach for Identifying Disease-Treatment Relations in Short Texts

Oana Frunza, Diana Inkpen, and Thomas Tran, *Member*, *IEEE*

Abstract—The Machine Learning (ML) field has gained its momentum in almost any domain of research and just recently has become a reliable tool in the medical domain. The empirical domain of automatic learning is used in tasks such as medical decision support, medical imaging, protein-protein interaction, extraction of medical knowledge, and for overall patient management care. ML is envisioned as a tool by which computer-based systems can be integrated in the healthcare field in order to get a better, more efficient medical care. This paper describes a ML-based methodology for building an application that is capable of identifying and disseminating healthcare information. It extracts sentences from published medical papers that mention diseases and treatments, and identifies semantic relations that exist between diseases and treatments. Our evaluation results for these tasks show that the proposed methodology obtains reliable outcomes that could be integrated in an application to be used in the medical care domain. The potential value of this paper stands in the ML settings that we propose and in the fact that we outperform previous results on the same data set.

Index Terms—Healthcare, machine learning, natural language processing.

1 Introduction

People care deeply about their health and want to be, now more than ever, in charge of their health and healthcare. Life is more hectic than has ever been, the medicine that is practiced today is an Evidence-Based Medicine (hereafter, EBM) in which medical expertise is not only based on years of practice but on the latest discoveries as well. Tools that can help us manage and better keep track of our health such as *Google Health*[1] *and Microsoft HealthVault*[2] are reasons and facts that make people more powerful when it comes to healthcare knowledge and management. The traditional healthcare system is also becoming one that embraces the Internet and the electronic world. Electronic Health Records (hereafter, EHR) are becoming the standard in the healthcare domain. Researches and studies show that the potential benefits of having an EHR system are[3]:

Health information recording and clinical data repositories—immediate access to patient diagnoses, allergies, and lab test results that enable better and time-efficient medical decisions;

Medication management—rapid access to information regarding potential adverse drug reactions, immunizations, supplies, etc;

Decision support—the ability to capture and use quality medical data for decisions in the workflow of healthcare; and

Obtain treatments that are tailored to specific health needs—rapid access to information that is focused on certain topics.

In order to embrace the views that the EHR system has, we need better, faster, and more reliable access to information. In the medical domain, the richest and most used source of information is Medline,[4] a database of extensive life science published articles. All research discoveries come and enter the repository at high rate (Hunter and Cohen [12]), making the process of identifying and disseminating reliable information a very difficult task. The work that we present in this paper is focused on two tasks: automatically identifying sentences published in medical abstracts (Medline) as containing or not information about diseases and treatments, and automatically identifying semantic relations that exist between diseases and treatments, as expressed in these texts. The second task is focused on three semantic relations: *Cure*, *Prevent*, and *Side Effect*.

The tasks that are addressed here are the foundation of an information technology framework that identifies and disseminates healthcare information. People want fast access to reliable information and in a manner that is suitable to their habits and workflow. Medical care related information (e.g., published articles, clinical trials, news, etc.) is a source of power for both healthcare providers and laypeople. Studies reveal that people are searching the web and read medical related information in order to be informed about their health. Ginsberg et al. [10] show how a new outbreak of the influenza virus can be detected from search engine query data.

Our objective for this work is to show what Natural Language Processing (NLP) and Machine Learning (ML) techniques—what representation of information and what classification algorithms—are suitable to use for identifying and classifying relevant medical information in short texts. We acknowledge the fact that tools capable of identifying reliable information in the medical domain stand as

1. https://www.google.com/health.
2. http://healthvault.com/.
3. http://healthcaretracker.wordpress.com/.
4. http://medline.cos.com/.

• *The authors are with the School of Information Technology and Engineering (SITE), University of Ottawa, 800 King Edward, Ottawa, Ontario K1N 6N5, Canada.*
E-mail: {ofrunza, diana, ttran}@site.uottawa.ca.

Manuscript received 13 Oct 2009; revised 20 Feb. 2010; accepted 21 Apr. 2010; published online 27 Aug. 2010.
Recommended for acceptance by E. Ferrari.
For information on obtaining reprints of this article, please send e-mail to: tkde@computer.org, and reference IEEECS Log Number TKDE-2009-10-0718.
Digital Object Identifier no. 10.1109/TKDE.2010.152.

building blocks for a healthcare system that is up-to-date with the latest discoveries. In this research, we focus on diseases and treatment information, and the relation that exists between these two entities. Our interests are inline with the tendency of having a personalized medicine, one in which each patient has its medical care tailored to its needs. It is not enough to read and know only about one study that states that a treatment is beneficial for a certain disease. Healthcare providers need to be up-to-date with all new discoveries about a certain treatment, in order to identify if it might have side effects for certain types of patients.

We envision the potential and value of the findings of our work as guidelines for the performance of a framework that is capable to find relevant information about diseases and treatments in a medical domain repository. The results that we obtained show that it is a realistic scenario to use NLP and ML techniques to build a tool, similar to an RSS feed, capable to identify and disseminate textual information related to diseases and treatments. Therefore, this study is aimed at designing and examining various representation techniques in combination with various learning methods to identify and extract biomedical relations from literature.

The contributions that we bring with our work stand in the fact that we present an extensive study of various ML algorithms and textual representations for classifying short medical texts and identifying semantic relations between two medical entities: diseases and treatments. From an ML point of view, we show that in short texts when identifying semantic relations between diseases and treatments a substantial improvement in results is obtained when using a hierarchical way of approaching the task (a pipeline of two tasks). It is better to identify and eliminate first the sentences that do not contain relevant information, and then classify the rest of the sentences by the relations of interest, instead of doing everything in one step by classifying sentences into one of the relations of interest plus the extra class of uninformative sentences.

The rest of the paper is organized as follows: Section 2 discusses related work, Section 3 presents our proposed approach to solve the task of identifying and disseminating healthcare information, Section 4 contains the evaluation and results obtained, Section 5 discussions, and Section 6 conclusions and future work.

2 Related Work

The most relevant related work is the work done by Rosario and Hearst [25]. The authors of this paper are the ones who created and distributed the data set used in our research. The data set consists of sentences from Medline[5] abstracts annotated with disease and treatment entities and with eight semantic relations between diseases and treatments. The main focus of their work is on entity recognition for diseases and treatments. The authors use Hidden Markov Models and maximum entropy models to perform both the task of entity recognition and the relation discrimination. Their representation techniques are based on words in context, part of speech information, phrases, and a medical lexical ontology—Mesh[6] terms. Compared to this work, our research is focused on different representation techniques, different classification models, and most importantly generates improved results with less annotated data.

The tasks addressed in our research are information extraction and relation extraction. From the wealth of research in these domains, we are going to mention some representative works. The task of relation extraction or relation identification is previously tackled in the medical literature, but with a focus on biomedical tasks: *subcellular-location* (Craven, [4]), *gene-disorder association* (Ray and Craven, [23]), and *diseases and drugs* (Srinivasan and Rindflesch, [26]). Usually, the data sets used in biomedical specific tasks use short texts, often sentences. This is the case of the first two related works mentioned above. The tasks often entail identification of relations between entities that co-occur in the same sentence.

There are three major approaches used in extracting relations between entities: co-occurrences analysis, rule-based approaches, and statistical methods. The co-occurrences methods are mostly based only on lexical knowledge and words in context, and even though they tend to obtain good levels of recall, their precision is low. Good representative examples of work on Medline abstracts include Jenssen et al. [14] and Stapley and Benoit [27].

In biomedical literature, rule-based approaches have been widely used for solving relation extraction tasks. The main sources of information used by this technique are either syntactic: part-of-speech (POS) and syntactic structures; or semantic information in the form of fixed patterns that contain words that trigger a certain relation. One of the drawbacks of using methods based on rules is that they tend to require more human-expert effort than data-driven methods (though human effort is needed in data-driven methods too, to label the data). The best rule-based systems are the ones that use rules constructed manually or semiautomatically—extracted automatically and refined manually. A positive aspect of rule-based systems is the fact that they obtain good precision results, while the recall levels tend to be low.

Syntactic rule-based relation extraction systems are complex systems based on additional tools used to assign POS tags or to extract syntactic parse trees. It is known that in the biomedical literature such tools are not yet at the state-of-the-art level as they are for general English texts, and therefore their performance on sentences is not always the best (Bunescu et al. [2]). Representative works on syntactic rule-based approaches for relation extraction in Medline abstracts and full-text articles are presented by Thomas et al. [28], Yakushiji et al. [29], and Leroy et al. [16]. Even though the syntactic information is the result of tools that are not 100 percent accurate, success stories with these types of systems have been encountered in the biomedical domain. The winner of the BioCreative II.5[7] task was a syntactic rule-based system, OpenDMAP described in Hunter et al. [13]. A good comparison of different syntactic

5. The sentences were extracted from the first 100 titles and the first 40 abstracts from each of the 59 files that are part of the Medline database from 2001.

6. http://www.nlm.nih.gov/mesh/meshhome.html.

7. http://www.biocreative.org/.

parsers and their contribution to extracting protein-protein interactions can be found in Miyao et al. [19].

The semantic rule-based approaches suffer from the fact that the lexicon changes from domain to domain, and new rules need to be created each time. Certain rules are created for biological corpora, medical corpora, pharmaceutical corpora, etc. Systems based on semantic rules applied to full-text articles are described by Friedman et al. [6], on sentences by Pustejovsky et al. [22], and on abstracts by Rindflesch et al. [24]. Some researchers combined syntactic and semantic rules from Medline abstracts in order to obtain better systems with the flexibility of the syntactic information and the good precision of the semantic rules, e.g., Gaizauskas et al. [8] and Novichkova et al. [20].

Statistical methods tend to be used to solve various NLP tasks when annotated corpora are available. Rules are automatically extracted by the learning algorithm when using statistical approaches to solve various tasks. In general, statistical techniques can perform well even with little training data. For extracting relations, the rules are used to determine if a textual input contains a relation or not. Taking a statistical approach to solve the relation extraction problem from abstracts, the most used representation technique is bag-of-words. It uses the words in context to create a feature vector (Donaldson et al. [5]) and (Mitsumori et al. [18]). Other researchers combined the bag-of-words features, extracted from sentences, with other sources of information like POS (Bunescu and Mooney [1]). Giuliano et al. [9] used two sources of information: sentences in which the relation appears and the local context of the entities, and showed that simple representation techniques bring good results.

Various learning algorithms have been used for the statistical learning approach with kernel methods being the popular ones applied to Medline abstracts (Li et al. [17]).

The task of identifying informative sentences is addressed in the literature mostly for the tasks of summarization and information extraction, and typically on such domains as newswire data, novels, medical, and biomedical domain. In the later mentioned domains, Goadrich et al. [11] used inductive logic techniques for information extraction from abstracts, while Ould et al. [21] experimented with bag-of-word features on sentences. Our work differs from the ones mentioned in this section by the fact that we combine different textual representation techniques for various ML algorithms.

3 The Proposed Approach

3.1 Tasks and Data Sets

The two tasks that are undertaken in this paper provide the basis for the design of an information technology framework that is capable to identify and disseminate healthcare information. The first task identifies and extracts informative sentences on diseases and treatments topics, while the second one performs a finer grained classification of these sentences according to the semantic relations that exists between diseases and treatments.

The problems addressed in this paper form the building blocks of a framework that can be used by healthcare providers (e.g., private clinics, hospitals, medical doctors, etc.), companies that build systematic reviews[8] (hereafter, SR), or laypeople who want to be in charge of their health by reading the latest life science published articles related to their interests. The final product can be envisioned as a browser plug-in or a desktop application that will automatically find and extract the latest medical discoveries related to disease-treatment relations and present them to the user. The product can be developed and sold by companies that do research in Healthcare Informatics, Natural Language Processing, and Machine Learning, and companies that develop tools like Microsoft Health Vault. The value of the product from an e-commerce point of view stands in the fact that it can be used in marketing strategies to show that the information that is presented is trustful (Medline articles) and that the results are the latest discoveries. For any type of business, the trust and interest of customers are the key success factors. Consumers are looking to buy or use products that satisfy their needs and gain their trust and confidence. Healthcare products are probably the most sensitive to the trust and confidence of consumers. Companies that want to sell information technology healthcare frameworks need to build tools that allow them to extract and mine automatically the wealth of published research. For example, in frameworks that make recommendations for drugs or treatments, these recommendations need to be based on acknowledged discoveries and published results, in order to gain the consumers' trust. The product value also stands in the fact that it can provide a *dynamic* content to the consumers, information tailored to a certain user (e.g., a set of diseases that the consumer is interested in).

The first task (task 1 or sentence selection) identifies sentences from Medline published abstracts that talk about diseases and treatments. The task is similar to a scan of sentences contained in the abstract of an article in order to present to the user-only sentences that are identified as containing relevant information (disease-treatment information).

The second task (task 2 or relation identification) has a deeper semantic dimension and it is focused on identifying disease-treatment relations in the sentences already selected as being informative (e.g., task 1 is applied first). We focus on three relations: *Cure, Prevent,* and *Side Effect*, a subset of the eight relations that the corpus is annotated with. We decided to focus on these three relations because these are most represented in the corpus while for the other five, very few examples are available. Table 1 presents the original data set, the one used by Rosario and Hearst [25], that we also use in our research. The numbers in parentheses represent the training and test set size. For example, for *Cure* relation, out of 810 sentences present in the data set, 648 are used for training and 162 for testing.

The approach used to solve the two proposed tasks is based on NLP and ML techniques. In a standard supervised ML setting, a training set and a test set are required. The training set is used to train the ML algorithm and the test set to test its performance. The objectives are to

8. Systematic reviews are summaries of research on a certain topic of interest. The topic can be a drug, disease, decision making step, etc.

TABLE 1
Data Set Description, Taken from Rosario and Hearst ('04)

Relationship	Definition and Example
Cure 810 (648, 162)	TREAT cures DIS *Intravenous immune globulin for recurrent spontaneous abortion*
Only DIS 616 (492, 124)	TREAT not mentioned *Social ties and susceptibility to the common cold*
Only TREAT 166 (132, 34)	DIS not mentioned *Flucticasome propionate is safe in recommended doses*
Prevent 63 (50, 13)	TREAT prevents the DIS *Statins for prevention of stroke*
Vague 36 (28, 8)	Very unclear relationship *Phenylbutazone and leukemia*
Side Effect 29 (24, 5)	DIS is a result of a TREAT *Malignant mesodermal mixed tumor of the uterus following irradiation*
NO Cure 4 (3, 1)	TREAT does not cure DIS *Evidence for double resistance to permethrin and malathion in head lice*
Total relevant: 1724 (1377, 347)	
Irrelevant 1771 (1416, 355)	Treat and DIS not present *Patients were followed up for 6 months*
Total: 3495 (2793, 702)	

In brackets, are the numbers of instances used for training and for testing, respectively.

build models that can later be deployed on other test sets with high performance.

For the work presented in this paper, the data sets contain sentences that are annotated with the appropriate information. Unlike in the work of Rosario and Hearst [25], in our research, the annotations of the data set are used to create a different task (task 1). It identifies informative sentences that contain information about diseases and treatments and semantic relations between them, versus noninformative sentences. This allows us to see how well NLP and ML techniques can cope with the task of identifying informative sentences, or in other words, how well they can weed out sentences that are not relevant to medical diseases and treatments.

Extracting informative sentences is a task by itself in the NLP and ML community. Research fields like summarization and information extraction are disciplines where the identification of informative text is a crucial task. The contributions and research value that are brought with this task stand in the usefulness of the results and the insights about the experimental settings for the task in the medical domain.

For the first task, the data sets are annotated with the following information: a label indicating that the sentence is informative, i.e., containing disease-treatment information, or a label indicating that the sentence is not informative. Table 2 gives an example of labeled sentences.

For the second task, the sentences have annotation information that states if the relation that exists in a sentence between the disease and treatment is *Cure*, *Prevent*, or *Side Effect*. These are the relations that are more represented in the original data set and also needed for our future research. We would like to focus on a few relations of interest and try to identify what predictive model and representation technique bring the best results. The task of identifying the three semantic relations is addressed in two ways:

Setting 1. Three models are built. Each model is focused on one relation and can distinguish sentences that contain the relation from sentences that do not. This setting is similar to a two-class classification task in which instances are labeled either with the relation in question (*Positive* label) or with nonrelevant information (*Negative* label);

Setting 2. One model is built, to distinguish the three relations in a three-class classification task where each sentence is labeled with one of the semantic relations.

Tables 3 and 4 present the data sets that we used for our two tasks.

In Table 4, the label "Positive" represents a sentence that contains the semantic relations (i.e., *Cure, Prevent,* or *Side Effect*), and "Negative" a sentence that does not contain information about any of the semantic relations but contains information about either a disease or a treatment labels *Treatment_Only*, *Disease_Only* in previous research by Rosario and Hearst [25].

Up to this point, we presented the two tasks separately as being two self-defined tasks since they can be used for other more complex tasks as well. From a methodological point of view, and, more importantly, from a practical point of view, they can be integrated together in a pipeline of tasks as a solution to a framework that is tailored to identify semantic relations in short texts and sentences, when it is not known a priori if the text contains useful information. The proposed pipeline solves task 1 first and then processes

TABLE 2
Examples of Annotated Sentences for the Sentence Selection Task

Label	Sentence
Informative sentence	*Urgent colonoscopy for the diagnosis and treatment of severe diverticular hemorrhage.*
Non-informative sentence	*In all cases a coproparasitological study was performed.*

TABLE 3
Data Sets Used for the First Task

	Informative sentences	Non-informative sentences
Training set	1225	1176
Test set	612	591

the results in task 2, so that in the end, only informative sentences are classified into the three semantic relations. The logic behind choosing to experiment with and report results for the pipeline of tasks is that we have to identify the best model that will get us closer to our main goal: being able to identify and classify reliably medical information. Using the pipeline of tasks, we eliminate some errors that can be introduced due to the fact that we would consider uninformative sentences as potential data when classifying sentences directly into semantic relations. We will show that the pipeline achieves much better results than a more straightforward approach of classifying in one step into one of the three relations of interest plus an extra class for uninformative sentences.

The pipeline is similar to a hierarchy of tasks in which the results of one task is given as input to the other. We believe that this can be a solution for identifying and disseminating relevant information tailored to a specific semantic relation because the second task is trying a finer grained classification of the sentences that already contain information about the relations of interest. This framework is appropriate for consumers that tend to be more interested in an end result that is more specific, e.g., relevant information only for the class *Cure*, rather than identifying sentences that have the potential to be informative for a wider variety of disease-treatment semantic relations.

3.2 Classification Algorithms and Data Representations

In ML, as a field of empirical studies, the acquired expertise and knowledge from previous research guide the way of solving new tasks. The models should be reliable at identifying informative sentences and discriminating disease-treatment semantic relations. The research experiments need to be guided such that high performance is obtained. The experimental settings are directed such that they are adapted to the domain of study (medical knowledge) and to the type of data we deal with (short texts or sentences), allowing for the methods to bring improved performance.

There are at least two challenges that can be encountered while working with ML techniques. One is to find the most suitable model for prediction. The ML field offers a suite of predictive models (algorithms) that can be used and deployed. The task of finding the suitable one relies heavily on empirical studies and knowledge expertise. The second one is to find a good data representation and to do feature engineering because features strongly influence the performance of the models. Identifying the right and sufficient features to represent the data for the predictive models, especially when the source of information is not large, as it is the case of sentences, is a crucial aspect that needs to be taken into consideration. These challenges are addressed by trying various predictive algorithms, and by using various textual representation techniques that we consider suitable for the task.

As classification algorithms, we use a set of six representative models: decision-based models (Decision trees), probabilistic models (Naïve Bayes (NB) and Complement Naïve Bayes (CNB), which is adapted for text with imbalanced class distribution), adaptive learning (AdaBoost), a linear classifier (support vector machine (SVM) with polynomial kernel), and a classifier that always predicts the majority class in the training data (used as a baseline). We decided to use these classifiers because they are representative for the learning algorithms in the literature and were shown to work well on both short and long texts. Decision trees are decision-based models similar to the rule-based models that are used in handcrafted systems, and are suitable for short texts. Probabilistic models, especially the ones based on the Naïve Bayes theory, are the state of the art in text classification and in almost any automatic text classification task. Adaptive learning algorithms are the ones that focus on hard-to-learn concepts, usually underrepresented in the data, a characteristic that appears in our short texts and imbalanced data sets. SVM-based models are acknowledged state-of-the-art classification techniques on text. All classifiers are part of a tool called Weka.[9] One can imagine the steps of processing the data (in our case textual information—sentences) for ML algorithms as the steps required to obtain a database table that contains as many columns as the number of features selected to represent the data, and as many rows as the number of data points from the collection (sentences in our case). The most difficult and important step is to identify

TABLE 4
Data Sets Used for the Second Task

	Training		Test	
	Positive	**Negative**	**Positive**	**Negative**
Cure	554	531	276	266
Prevent	42	531	21	266
SideEffect	20	531	10	266

9. http://www.cs.waikato.ac.nz/ml/weka/.

which features should be selected to represent the data. A special column in this table represents the label of each instance. An instance represents a row that contains values for the selected features. The ML algorithms that are using this data representation to create predictive models should capture correlations between features, feature values, and labels, in order to obtain good prediction labels on future test data. The innovation and contribution that immerge form these experimental settings stand in identifying the most informative features for the task to feed the models, while using a suitable predictive algorithm in order to increase the chance of predicting correct labels for new texts processed in the future. The following sections present the data representation techniques.

3.2.1 Bag-of-Words Representation

The bag-of-words (BOW) representation is commonly used for text classification tasks. It is a representation in which features are chosen among the words that are present in the training data. Selection techniques are used in order to identify the most suitable words as features. After the feature space is identified, each training and test instance is mapped to this feature representation by giving values to each feature for a certain instance. Two most common feature value representations for BOW representation are: binary feature values—the value of a feature can be either 0 or 1, where 1 represents the fact that the feature is present in the instance and 0 otherwise; or frequency feature values—the value of the feature is the number of times it appears in an instance, or 0 if it did not appear.

Because we deal with short texts with an average of 20 words per sentence, the difference between a binary value representation and a frequency value representation is not large. In our case, we chose a frequency value representation. This has the advantage that if a feature appears more than once in a sentence, this means that it is important and the frequency value representation will capture this—the feature's value will be greater than that of other features. The selected features are words delimited by spaces and simple punctuation marks such as *(,) , [,] ,. ,'*. We keep only the words that appeared at least three times in the training collection, contain at least one alphanumeric character, are not part of an English list of stop words,[10] and are longer than three characters. The frequency threshold of three is commonly used for text collections because it removes noninformative features and also strings of characters that might be the result of a wrong tokenization when splitting the text into words. Words that have length of two or one character are not considered as features because of two other reasons: possible incorrect tokenization and problems with very short acronyms in the medical domain that could be highly ambiguous (could be an acronym or an abbreviation of a common word).

3.2.2 NLP and Biomedical Concepts Representation

The second type of representation is based on syntactic information: noun-phrases, verb-phrases, and biomedical concepts identified in the sentences. In order to extract this type of information, we used the Genia[11] tagger tool. The tagger analyzes English sentences and outputs the base forms, part-of-speech tags, chunk tags, and named entity tags. The tagger is specifically tuned for biomedical text such as Medline abstracts. Fig. 1 presents an example of the output of the Genia tagger for the sentence: "*Inhibition of NF-kappaB activation reversed the anti-apoptotic effect of isochamaejasmin.*" The noun and verb-phrases identified by the tagger are features used for the second representation technique.

Inhibition	Inhibition	NN	B-NP	O
of	of	IN	B-PP	O
NF-kappaB	NF-kappaB	NN	B-NP	B-protein
activation	activation	NN	I-NP	O
reversed	reverse	VBD	B-VP	O
the	the	DT	B-NP	O
anti-apoptotic	anti-apoptotic	JJ	I-NP	O
effect	effect	NN	I-NP	O
of	of	IN	B-PP	O
isochamaejasmin	isochamaejasmin	NN	B-NP	O
.	.	.	O	O

Fig. 1. Example of Genia tagger output including for each word: its base form, its part-of-speech, beginning (B), inside (I), outside (O) tags for the word, and the final tag for the phrase.

We ran the Genia tagger on the entire data set. We extracted only noun-phrases, verb-phrases, and biomedical concepts as potential features from the output of each sentence present in the data set.

The following preprocessing steps are applied in order to identify the final set of features to be used for classification: removing features that contain only punctuation, removing stop words (using the same list of words as for our BOW representation), and considering valid features only the lemma-based forms. We chose to use lemmas because there are a lot of inflected forms (e.g., plural forms) for the same word and the lemmatized form (the base form of a word) will give us the same base form for all of them. Another reason is to reduce the data sparseness problem. Dealing with short texts, very few features are represented in each instance; using lemma forms alleviates this problem. Experiments are performed when using as features only the final set of identified noun-phrases, only verb-phrases, only biomedical entities, and with combinations of all these features. When combining the features, the feature vector for each instance is a concatenation of all features.

3.2.3 Medical Concepts (UMLS) Representation

In order to work with a representation that provides features that are more general than the words in the abstracts (used in the BOW representation), we also used the Unified Medical Language system[12] (hereafter, UMLS) concept representations. UMLS is a knowledge source developed at the US National Library of Medicine (hereafter, NLM) and it contains a metathesaurus, a semantic network, and the specialist lexicon for biomedical domain. The metathesaurus is organized around

10. http://www.site.uottawa.ca/~diana/csi5180/StopWords. Stop words are function words that appear in every document (e.g., *the, it, of,* and *an*) and therefore do not help in classification.

11. http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/tagger/.

12. http://www.nlm.nih.gov/pubs/factsheets/umls.html.

Meta Candidates (6)
861 Risk [Qualitative Concept, Quantitative Concept]
694 Increased (Increased (qualifier value)) [Functional Concept]
623 Increase (Increase (qualifier value)) [Functional Concept]
601 Acquired (Acquired (qualifier value)) [Temporal Concept]
601 Obtained (Obtained (attribute)) [Functional Concept]
588 Increasing (Increasing (qualifier value)) [Functional Concept]

Fig. 2. Example of MetaMap system output.

concepts and meanings; it links alternative names and views of the same concept and identifies useful relationships between different concepts.

UMLS contains over 1 million medical concepts, and over 5 million concept names which are hierarchical organized. All concepts are assigned at least one semantic type from the semantic network providing a generalization of the existing relations between concepts. There are 135 semantic types in the knowledge base that are linked through 54 relationships.

In addition to the UMLS knowledge base, NLM created a set of tools that allow easier access to the useful information. MetaMap[13] is a tool created by NLM that maps free text to medical concepts in the UMLS, or equivalently, it discovers metathesaurus concepts in text. With this software, text is processed through a series of modules that in the end will give a ranked list of all possible concept candidates for a particular noun-phrase. For each of the noun-phrases that the system finds in the text, variant noun-phrases are generated. For each of the variant noun-phrases, candidate concepts (concepts that contain the noun-phrase variant) from the UMLS metathesaurus are retrieved and evaluated. The retrieved concepts are compared to the actual phrase using a fit function that measures the text overlap between the actual phrase and the candidate concept (it returns a numerical value). The best of the candidates are then organized according to the decreasing value of the fit function. We used the top concept candidate for each identified phrase in an abstract as a feature.

Fig. 2 presents an example of the output of the MetaMap system for the phrase "*to an increased risk*." The information present in the brackets, "Qualitative Concept, Quantitative Concept" for the candidate with the fit function value 861 is the concept used as feature in the UMLS representation. Another reason to use a UMLS concept representation is the ***concept drift*** phenomenon that can appear in a BOW representation. Especially in the medical domain texts, this is a frequent problem as stated by Cohen et al. [3]: new articles that publish new research on a certain topic bring with them new terms that might not match the ones that were seen in the training process in a certain moment of time. The UMLS concepts also help with the data sparseness problem and give a better coverage of the features in each sentence instance.

Experiments for the two tasks tackled in this research were performed with each individual above-mentioned representations, plus their combinations. We combined the BOW, UMLS, NLP, and biomedical concepts, by putting the features together to represent an instance.

13. http://mmtx.nlm.nih.gov/.

4 Evaluation and Results

This section discusses the evaluation measures and presents the results of the two tasks using the methodology described above.

4.1 Evaluation Measures

The most common used evaluation measures in the ML settings are: accuracy, precision, recall, and F-measure. All these measures are computed form a confusion matrix (Kohavi and Provost [15]) that contains information about the actual classes, the true classes and the classes predicted by the classifier. The test set on which the models are evaluated contain the true classes and the evaluation tries to identify how many of the true classes were predicted by the model classifier. In the ML settings, special attention needs to be directed to the evaluation measures that are used. For data sets that are highly imbalanced (one class is overrepresented in comparison with another), standard evaluation measures like accuracy are not suitable. Because our data sets are imbalanced, we chose to report in addition to accuracy, the macroaveraged F-measure. We decided to report macro and not microaveraged F-measure because the macromeasure is not influenced by the majority class, as the micromeasure is. The macromeasure better focuses on the performance the classifier has on the minority classes. The formulas for the evaluation measures are: $\boldsymbol{Accuracy} =$ the total number of correctly classified instances; $\boldsymbol{Recall} =$ the ratio of correctly classified positive instances to the total number of positives. This evaluation measure is known to the medical research community as sensitivity. $\boldsymbol{Precision} =$ the ratio of correctly classified positive instances to the total number of classified as positive. $\boldsymbol{F\text{-}measure} =$ the harmonic mean between precision and recall.

4.2 Results for the Task of Identifying Informative Sentences (Task 1)

This section presents the results for the first task, the one of identifying whether sentences are informative, i.e., containing information about diseases and treatments, or not. The ML settings are created for a two-class classification task and the representations are the ones mentioned in the previous section, while the baseline on which we need to improve is given by the results of a classifier that always predicts the majority class.

Fig. 3 presents the results obtained when using as representation features verb-phrases identified by the Genia tagger. When using this representation, the results are close to baseline. The reason why this happens for all algorithms that we use is the fact that the texts are short and the selected features are not well represented in an instance. We have a data sparseness problem: it is the case when a lot of features have value 0 for a particular instance.

Fig. 4 presents the results obtained using as representation features noun-phrases selected by the Genia tagger. Compared to previous results, we can observe a slight improvement in both accuracy and F-measure. The best results are obtained by the CNB classifier. We believe that the slight improvement is due to a reduction of the sparseness problem: noun-phrases are more frequently

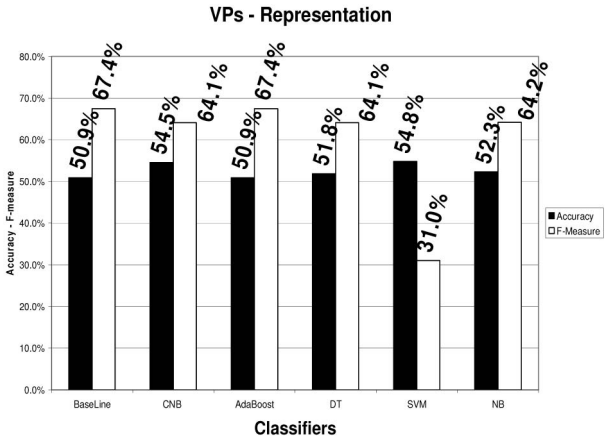

Fig. 3. Accuracy and F-measure results when using verb-phrases as features for task 1.

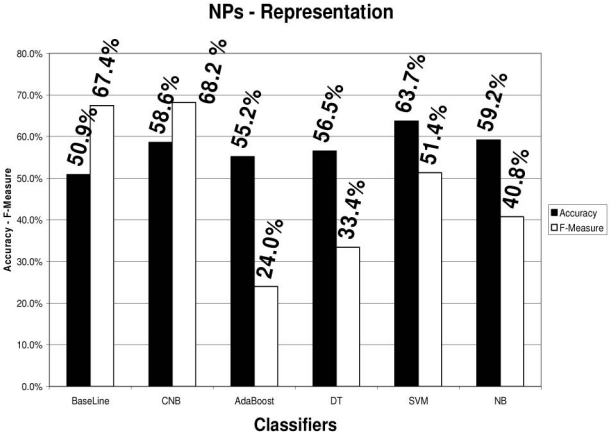

Fig. 4. Accuracy and F-measure results when using noun-phrases as features for task 1.

present in short texts than verb-phrases. Fig. 5 presents the best results obtained so far. An increase of almost 5 percentage points, for both accuracy and F-measure is obtained when using as representation features biomedical entities extracted by the Genia tagger and CNB as classifier. An increase in results for the other classifiers can be also observed.

This increase can be caused by the fact that, when present in sentences, the biomedical entities have a stronger predicting value. The entities identify better if a sentence is informative or not and this is something that we would hope to happen. If a sentence contains a good proportion of biomedical entities this should trigger a higher chance of labeling a sentence as informative.

Fig. 6 presents accuracy and F-measure results for all classifiers when noun-phrases, identified by the Genia tagger and biomedical entities are used as representation features. Compared to previous representation techniques, the results presented in Fig. 6 follow what will become a trend, an increase in results when more informative and diverse representation techniques are used. With a representation that combines noun-phrases and biomedical entities and CNB as classifier, an increase of 2 percentage points is obtained compared to the results when only biomedical concepts are used (Fig. 5). An increase of 8 percentage points

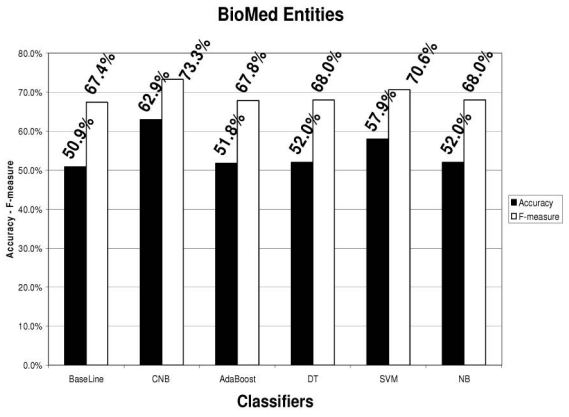

Fig. 5. Accuracy and F-measure results when using biomedical concepts as features, task 1.

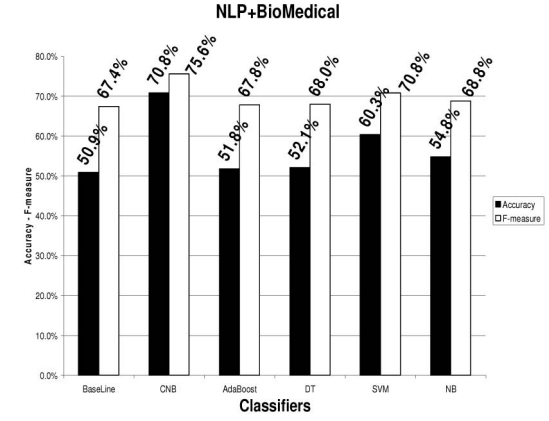

Fig. 6. Accuracy and F-measure results when using NLP and biomedical concepts as features, task 1.

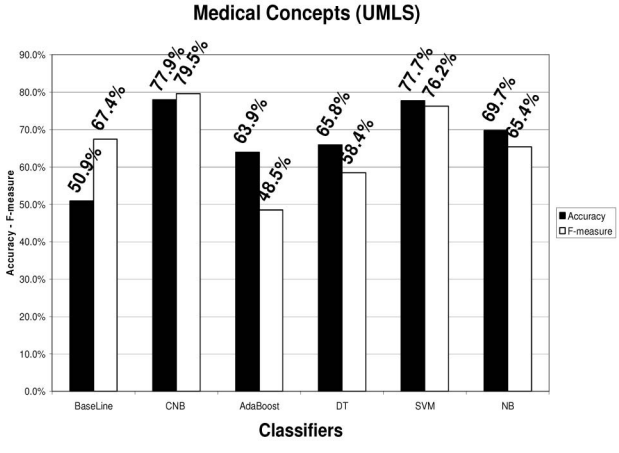

Fig. 7. Accuracy and F-measure results when using UMLS concepts as features for task 1.

is obtained compared to only a noun-phrase representation (Fig. 4).

In Fig. 7, we use as representation technique UMLS concepts—medical domain-specific concepts identified by the MetaMap tool. Compared to all previous results, this representation technique for the CNB classifier obtains the best results so far, with an increase of almost 4 percentage points. We believe that this is because these concepts are

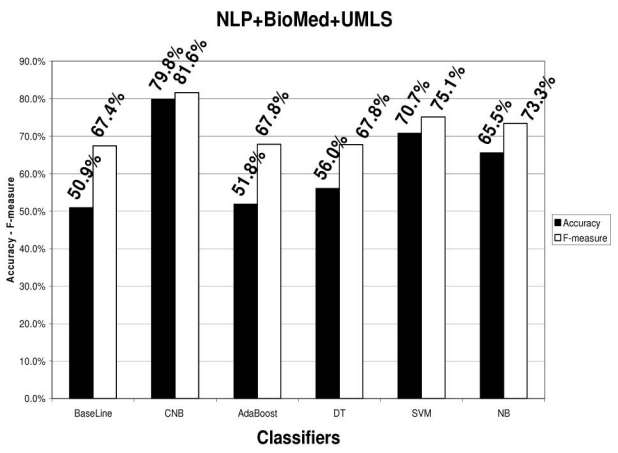

Fig. 8. Accuracy and F-measure results when using NLP, biomedical, and UMLS concepts as features, task 1.

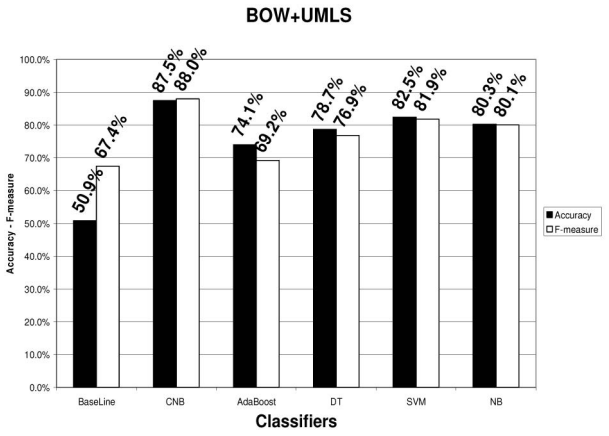

Fig. 10. Accuracy and F-measure results when using BOW and UMLS concepts as features for task 1.

more general than the biomedical concepts and the feature space is not that sparse. Classifiers tend to obtain better results when the feature space is well represented by an instance. This observation is what we believe happened with the results presented in Fig. 7.

A representation based on noun-phrases, verb-phrases, biomedical concepts, and UMLS concepts brings again an increase in results compared to the ones presented in Fig. 7. This increase of 2 percentage points can be observed in Fig. 8. As stated before, the trend of increase in results is due to the representation that captures more information.

The bag-of-words representation technique is known in the literature to be one that is hard to beat. Even though is not a very sophisticated method—it contains only the words in context; it is one that often obtains good results. In our experiments, the BOW representation (Fig. 9) obtains the best results between all the representation techniques mentioned in this section so far. An increase of almost 2 percentage points is obtained compared to the results in Fig. 8.

In Fig. 10, we present the results for a BOW plus UMLS concepts representation. Even though the BOW representation is one that gives good results, when used in combinations with other types of representations, the performance can be improved. This observation can be drawn from Fig. 10 where a 2 percentage points improvement is obtained for CNB, compared to the BOW representation; an improvement of 10 percentage points is obtained when comparing to the UMLS representation. For all the other classifiers, an increase in results can be observed as well. The results obtained by only a BOW representation can be further improved when these features are combined with the noun-phrases, verb-phrases, and biomedical concepts. Fig. 11 presents results for this representation technique with all the classification algorithms. The best result is improved with 1 percentage point compared to the one in Fig. 10.

For this current task, identifying which sentences from the abstracts of Medline articles that contain informative sentences for diseases and treatments, the best results obtained are the one presented in Fig. 12. The representation technique that uses BOW features, UMLS concepts, noun and verb-phrases, and biomedical concepts with the CNB classifier obtain a 90.72 percent F-measure and 90.36 percent accuracy. These increases in results are due to the fact that all these various types of features create a rich and predictive feature space for the classifiers.

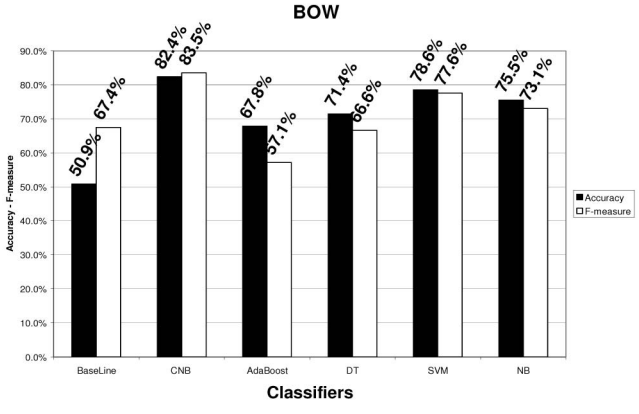

Fig. 9. Accuracy and F-measure results when using BOW features for task 1.

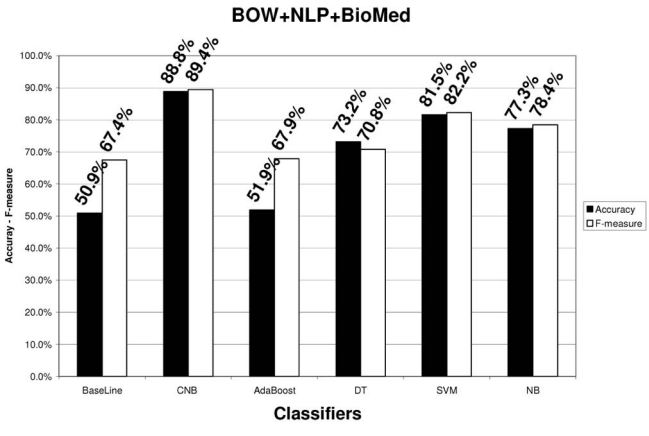

Fig. 11. Accuracy and F-measure results when using BOW, NLP, and biomedical features, task 1.

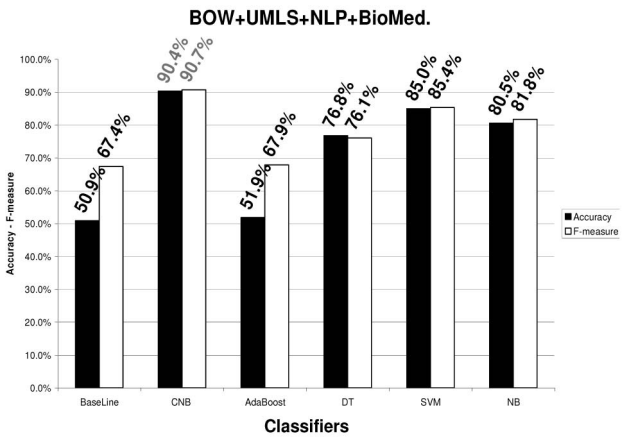

Fig. 12. Accuracy and F-measure results when using BOW, NLP, biomedical, and UMLS concepts features, task 1.

Even if some of the features alone are not the ones that obtain good performance, e.g., verb-phrases, when combined with other types of features form a representation model capable to predict with 90 percent accuracy if a sentence is informative or not. The fact that the best results for this task are obtained when all features are put together shows that a good representation technique is a crucial aspect of a classification task. In order to statistically support our conclusions, we run t-tests at 95 percent confidence levels for the CNB classifier with all the representation techniques for both accuracy and F-measure. Even though some of the differences are small, they are significantly different, except the one between Figs. 8 and 9. Based on our experiments, we can conclude and suggest as future guidelines for similar tasks that the richer and more informative the representation technique is, the better the performance results. For a good performance level, we suggest a combination of all the features.

4.3 Results for the Task of Identifying Semantic Relations (Task 2)

The focus for the second task is to automatically identify which sentences contain information for the three semantic relations: *Cure, Prevent*, and *Side Effect*. The reported results are based on similar settings to the ones used for the previous task. Since imbalanced data sets are used for this task, the evaluation measure that we are going to report is the F-measure. Due to space issues, we are going to present the best results obtained for all settings. The best results are chosen from all the representation techniques and all classification algorithms that we also used for the first task. The labels on the x-axis stand for the name of the semantic relation, the representation technique, and the classification algorithm used.

In Fig. 13, we present the results when using Setting 1, described in Section 3.1, as the setup for the experiment. On the x-axis, we present for each relation the best F-measure result, the representation technique, and the classifier that obtained the result. For example, for the *Cure* relation, the combination of BOW features, noun-phrases and verb-phrases, biomedical and UMLS concepts, with SVM as a classifier, obtained the 87.10 percent result for F-measure. SVM and NB with rich feature representations are the

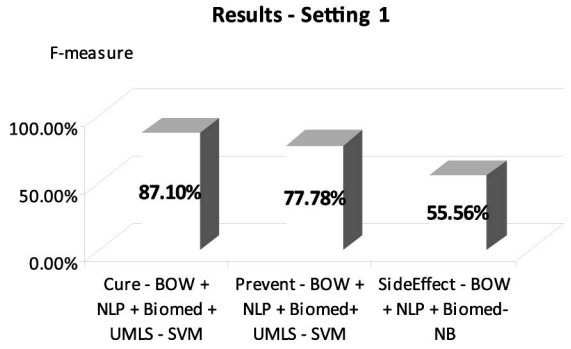

Fig. 13. Results for the second task, Setting 1.

setups that obtained the best results. Fig. 14 presents the best results that we obtain for the second task, a level of almost 100 percent F-measure for the *Cure* relation, 100 percent F-measure for *Prevent* relation, and 75 percent F-measure for *Side Effect*. For this setting, we train a model for all three relations in the same time, and we distinguish sentences between these three relations.

For this setting, the NB classifier with combinations of various representation features is the one that obtains the best results for all relations. The improvement over the other settings can be due to the fact that the combination of classifier and features has a good predicting value for a model trained on the three relations. Each of the relations can be well-defined and predicted when using the model that we propose in Setting 2. The fact that we achieve close to perfection prediction suggests that the choice of classifier and representation technique are key factors for a supervised classification task, even for semantically charged tasks like ours. The good performance results that we obtain with the second setting also suggest that a prior triage of sentences, informative versus noninformative can be crucial for a finer grained classification of relations between entities. Setting 1 uses all the sentences, including those that do not contain information about the three relations of interests, while in Setting 2, we used as training data only sentences that we knew a priori to contain one of the three relations. This observation for the results of the second setting also validates our choice of proposing the first task, identify which sentences are informative and which not. For good performance level in the relation classification task, we need to weed out noninformative sentences.

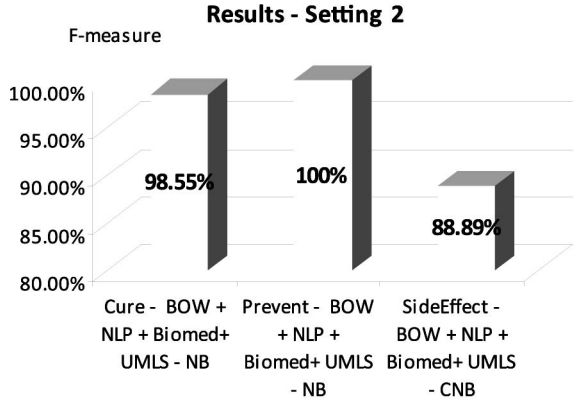

Fig. 14. Results for the second task, Setting 2.

TABLE 5
F-Measure Results for the Pipeline-Task 1 Followed by Task 2

Semantic Relation	F-measure Task1	F-measure Task2	F-measure Pipeline
Cure	90.72%	98.55%	**89.40%**
Prevent	90.72%	100%	**90.72%**
SideEffect	90.72%	88.89%	**80.64%**

4.4 Results for the Pipeline—Task 1 Followed by Task 2

In this section, we present the evaluation results for the pipeline of the two tasks. When looking at the results that we obtain for the second task, the best setting was the one in which we classify sentences already known to be informative (Setting 2). This observation let us believe that a pipeline of the two tasks is a viable solution for our goal.

To show that a pipeline of results is better as a solution for identifying semantic relations in informative sentences, we need to compare its results to the results of a model that classifies sentences into four-classes directly: the three semantic relations *Cure*, *Prevent*, *SideEffect* and the class for sentences that are uninformative.

The results for the pipeline of tasks are obtained by multiplying the evaluation measures acquired by the first task with the evaluation measure for the second task for each semantic relation. To be consistent, we report the F-measure results. For the first task, the best F-measure result of 90.72 percent is obtained by the CNB classifier using a combination of all types of features (Fig. 12). For the second task, the best F-measure results are obtained by the NB classifier using a combination of all types of features for all three semantic relations (Fig. 14). Table 5 presents the results for the pipeline of tasks.

The results for a scenario in which we solve the task of identifying sentences relevant to one of the three semantic relations by deploying a four-class trained classifier are presented in Fig. 15. In this experiment, we used all classifiers and all representation techniques that we propose in this paper on a data set that consists of sentences that are either labeled as uninformative, sentences from task 1, or with one of the three semantic relations, the data set used in task 2. To be consistent with all other experiments, we report F-measure results as well. Since reporting the results of every setting would take a lot of space, we decided to report only the best ones. Fig. 15 presents these results. The representation and classification algorithms mentioned in the legend correspond to the left-to-right results. The pipeline of tasks clearly outperforms the four-class classification scenario for the *Prevent* and *SideEffect* class. An increase of 30 percentage points is obtained for the *Prevent* class and 18 percentage points for the *SideEffect* class. For the class *Cure*, the four-class methodology was superior with 4 percentage points. The reason for this could be the fact that the *Cure* class is well represented in the data set and in the end it has a higher chance to be correctly classified in almost any ML scenario. The important type of error that can occur in a four-way classification task is the high rate of false negatives for the three classes of interest. Some sentences that belong to one of the classes of interest get classified in the fourth class that contains uninformative sentences. In a pipeline of tasks scenario, if the result of the first task is high and uninformative sentences are removed, then informative sentences have higher changes to be classified into the right class, especially when the choices are reduced by one.

Usually the amount of data that is classified as being nonrelevant is higher than the one relevant to a particular relation. In a four-way classification task, the majority class overwhelms the underrepresented ones while in a pipeline of tasks the balance between that relevant data and nonrelevant one is higher and the classifiers have better chances of distinguish between them.

The fact that for the two underrepresented classes, we obtain a high increase in results suggests that a pipeline of tasks is superior in performance to a four-class classification task.

5 Discussion

This section discusses the results we obtained for the two tasks in this study. For the first task, the one of identifying informative sentences, the results show that probabilistic models based on Naïve Bayes formula, obtain good results. The fact that the SVM classifier performs well shows that the current discoveries are in line with the literature. These two classifiers have always been shown to perform well on text classification tasks. Even though the independence of features is violated when using Naïve Bayes classifiers, they still perform very well. The AdaBoost classifier was outperformed by the other classifiers, which is a little

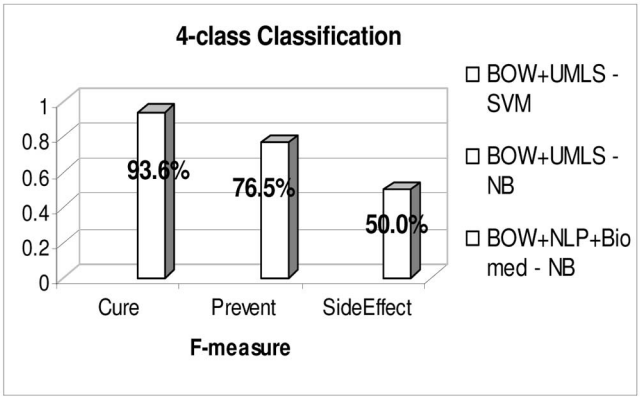

Fig. 15. F-measure results for four-class classification.

surprising taking into account the fact that it is designed to focus on hard-to-learn concepts. In our previous experience, it was shown to perform well on medical domain texts with imbalanced classes (Frunza and Inkpen [7]). One reason why the AdaBoost classifier did not perform well might be that fact that in previous experiments we used the entire abstract as source of information while in this current study we use sentences.

In NLP and ML community, BOW is a representation technique that even though it is simplistic, most of the times it is really hard to outperform. As shown in Fig. 9, the results obtained with this representation are among the best one, but for both tasks, we outperform it when we combine it with more structured information such as medical and biomedical concepts.

One of the major contributions of this work is the fact that the current experiments show that additional information in the representation settings brings improvements for the task of identifying informative sentences. The task itself is a knowledge-charged task; the labeling process involves a human-intensive annotation process since relations between entities need to be manually identified. The experiments designed for the automatic task aim to show that classifiers perform better when richer information is provided. In the first task, the CNB classifier using all the representation techniques obtains the best result of 90 percent F-measure which is statistically significant. The classifier is specially designed for imbalanced data, the fact that it proved to be one of the best in text classification tasks, even on short texts, was somewhat a foreseeable result.

The results obtained for the second task suggest that when the focus of a task is to obtain good reliable results, extra analysis is required. The best results are obtained with Setting 2 when a model is built and trained on a data set that contains all three data sets for the three relations. The representation and the classification algorithms were able to make the distinction between the relations and to obtain the best results for this task. Similar observations as the ones obtained for the first task are valid: probabilistic models combined with more informative feature representation bring the best results. The best results obtained are: 98 percent F-measure for the class *Cure*, 100 percent F-measure for the class *Prevent*, and 75 percent F-measure for the *SideEffect* class.

The fact that we obtain the best results when using Setting 2 also validates our proposed methodology for a pipeline of tasks in order to better classify relevant information in the three semantic relations. It is more efficient to solve the second task when using data that are known a priori to contain information about the relations in question, rather than identifying which sentences are uninformative as well.

In order to better validate the choices made in terms of representation and classification algorithms and to directly compare with the previous work, additional experiments for all eight semantic relations originally annotated on the data set were performed. These experiments are addressing exactly the same task as the previous work (Rosario and Hearst [25]) and are evaluated with the same evaluation measure, accuracy. Due to space constrains, we will report in Fig. 16 only the best results with the algorithms and representations that we used for this task.

The first bars of results are obtained with the best model for each of the eight relations (e.g., for *Cure,* the representation that obtains the best results is reported, a representation that can be different from the one for another relation; the label of each set of bars describes the representation); the second bars of results report the model that obtains the best accuracy over all relations (one representation and one classification algorithm are reported for all relations—CNB with BOW+NLP+Biomed features), and the third bars of results represent the previous results obtained by Rosario and Hearst [25].

The accuracy measure is reported since it is the measure that was reported in the previous work. As depicted in the figure, the results obtained in this study outperform the previous ones. In one case, the same low results are obtained; for example, for the *No_Cure* class, the low results are due to the fact that this class is underrepresented in the data set, by only four examples in total.

The class *Vague* obtains similar results when one model is used for all relations, but it outperforms previous results when the best model is chosen for this class. For the other relation, our results are better with either the same model for all relations, or for the best one for each.

6 Conclusions and Future Work

The conclusions of our study suggest that domain-specific knowledge improves the results. Probabilistic models are stable and reliable for tasks performed on short texts in the medical domain. The representation techniques influence the results of the ML algorithms, but more informative representations are the ones that consistently obtain the best results.

The first task that we tackle in this paper is a task that has applications in information retrieval, information extraction, and text summarization. We identify potential improvements in results when more information is brought in the representation technique for the task of classifying short medical texts. We show that the simple BOW approach, well known to give reliable results on text classification tasks, can be significantly outperformed when adding more complex and structured information from various ontologies.

The second task that we address can be viewed as a task that could benefit from solving the first task first. In this study, we have focused on three semantic relations between diseases and treatments. Our work shows that the best results are obtained when the classifier is not overwhelmed by sentences that are not related to the task. Also, to perform a triage of the sentences (task 1) for a relation classification task is an important step. In Setting 1, we included the sentences that did not contain any of the three relations in question and the results were lower than the one when we used models trained only on sentences containing the three relations of interest. These discoveries validate the fact that it is crucial to have the first step to weed out uninformative sentences, before looking deeper into classifying them. Similar findings and conclusions can be made for the representation and classification techniques for task 2.

The above observations support the pipeline of tasks that we propose in this work. The improvement in results of 14 and 18 percentage points that we obtain for two of the classes in question shows that a framework in which tasks 1

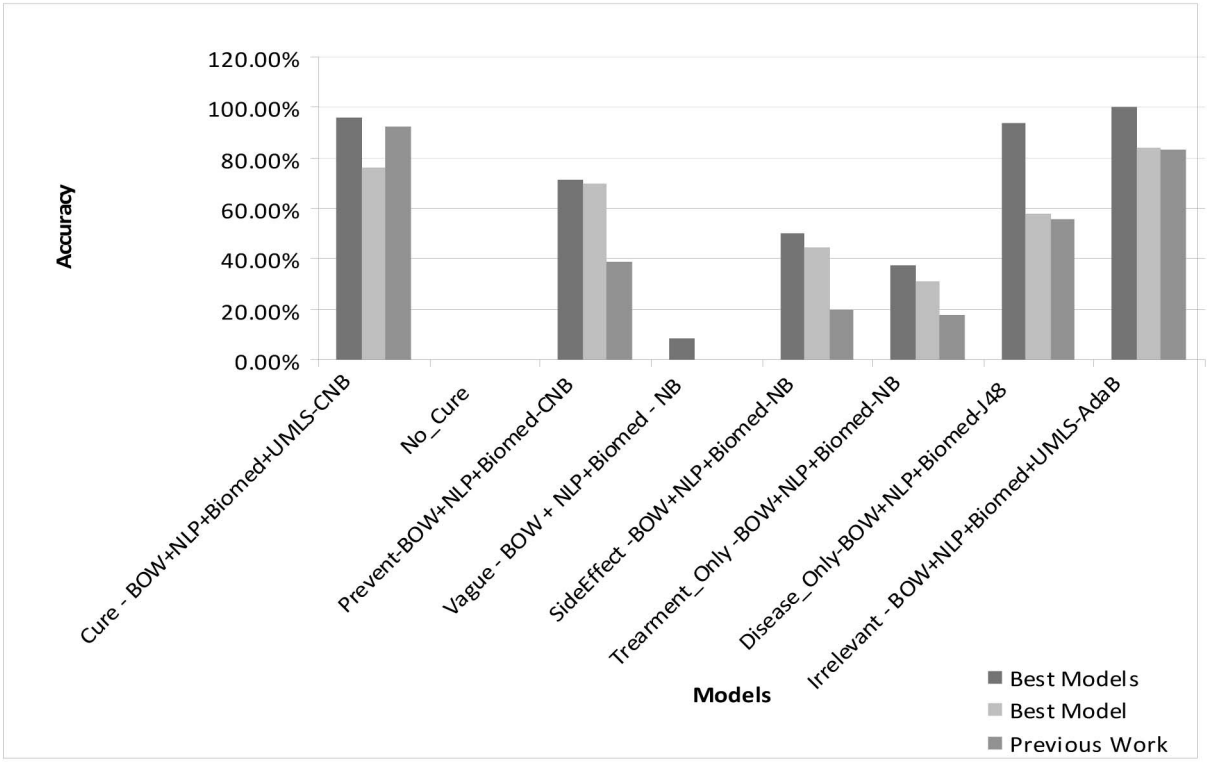

Fig. 16. Results for all annotated relations in the data set.

and 2 are used in pipeline is superior to when the two tasks are solved in one step by a four-way classification.

Probabilistic models combined with a rich representation technique bring the best results.

As future work, we would like to extend the experimental methodology when the first setting is applied for the second task, to use additional sources of information as representation techniques, and to focus more on ways to integrate the research discoveries in a framework to be deployed to consumers. In addition to more methodological settings in which we try to find the potential value of other types of representations, we would like to focus on source data that comes from the web. Identifying and classifying medical-related information on the web is a challenge that can bring valuable information to the research community and also to the end user. We also consider as potential future work ways in which the framework's capabilities can be used in a commercial recommender system and in integration in a new EHR system. Amazon representative Jeff Bezos said: "Our experience with user interfaces and high-performance computing are ideally suited to help healthcare. We nudge people's decision making and behavior with the gentle push of data *[...]*".[14]

REFERENCES

[1] R. Bunescu and R. Mooney, "A Shortest Path Dependency Kernel for Relation Extraction," *Proc. Conf. Human Language Technology and Empirical Methods in Natural Language Processing (HLT/EMNLP)*, pp. 724-731, 2005.

[2] R. Bunescu, R. Mooney, Y. Weiss, B. Schölkopf, and J. Platt, "Subsequence Kernels for Relation Extraction," *Advances in Neural Information Processing Systems*, vol. 18, pp. 171-178, 2006.

[3] A.M. Cohen and W.R. Hersh, and R.T. Bhupatiraju, "Feature Generation, Feature Selection, Classifiers, and Conceptual Drift for Biomedical Document Triage," *Proc. 13th Text Retrieval Conf. (TREC)*, 2004.

[4] M. Craven, "Learning to Extract Relations from Medline," *Proc. Assoc. for the Advancement of Artificial Intelligence*, 1999.

[5] I. Donaldson et al., "PreBIND and Textomy: Mining the Biomedical Literature for Protein-Protein Interactions Using a Support Vector Machine," *BMC Bioinformatics*, vol. 4, 2003.

[6] C. Friedman, P. Kra, H. Yu, M. Krauthammer, and A. Rzhetsky, "GENIES: A Natural Language Processing System for the Extraction of Molecular Pathways from Journal Articles," *Bioinformatics*, vol. 17, pp. S74-S82, 2001.

[7] O. Frunza and D. Inkpen, "Textual Information in Predicting Functional Properties of the Genes," *Proc. Workshop Current Trends in Biomedical Natural Language Processing (BioNLP) in conjunction with Assoc. for Computational Linguistics (ACL '08)*, 2008.

[8] R. Gaizauskas, G. Demetriou, P.J. Artymiuk, and P. Willett, "Protein Structures and Information Extraction from Biological Texts: The PASTA System," *Bioinformatics*, vol. 19, no. 1, pp. 135-143, 2003.

[9] C. Giuliano, L. Alberto, and R. Lorenza, "Exploiting Shallow Linguistic Information for Relation Extraction from Biomedical Literature," *Proc. 11th Conf. European Chapter of the Assoc. for Computational Linguistics*, 2006.

[10] J. Ginsberg, H. Mohebbi Matthew, S.P. Rajan, B. Lynnette, S.S. Mark, and L. Brilliant, "Detecting Influenza Epidemics Using Search Engine Query Data," *Nature*, vol. 457, pp. 1012-1014, Feb. 2009.

[11] M. Goadrich, L. Oliphant, and J. Shavlik, "Learning Ensembles of First-Order Clauses for Recall-Precision Curves: A Case Study in Biomedical Information Extraction," *Proc. 14th Int'l Conf. Inductive Logic Programming*, 2004.

[12] L. Hunter and K.B. Cohen, "Biomedical Language Processing: What's beyond PubMed?" *Molecular Cell*, vol. 21-5, pp. 589-594, 2006.

14. http://www.healthcare-informatics.com.

[13] L. Hunter, Z. Lu, J. Firby, W.A. Baumgartner Jr., H.L. Johnson, P.V. Ogren, and K.B. Cohen, "OpenDMAP: An Open Source, Ontology-Driven Concept Analysis Engine, with Applications to Capturing Knowledge Regarding Protein Transport, Protein Interactions and Cell-Type-Specific Gene Expression," *BMC Bioinformatics,* vol. 9, article no. 78, Jan. 2008.
[14] T.K. Jenssen, A. Laegreid, J. Komorowski, and E. Hovig, "A Literature Network of Human Genes for High-Throughput Analysis of Gene Expression," *Nature Genetics,* vol. 28, no. 1, pp. 21-28, 2001.
[15] R. Kohavi and F. Provost, "Glossary of Terms," *Machine Learning,* Editorial for the Special Issue on Applications of Machine Learning and the Knowledge Discovery Process, vol. 30, pp. 271-274, 1998.
[16] G. Leroy, H.C. Chen, and J.D. Martinez, "A Shallow Parser Based on Closed-Class Words to Capture Relations in Biomedical Text," *J. Biomedical Informatics,* vol. 36, no. 3, pp. 145-158, 2003.
[17] J. Li, Z. Zhang, X. Li, and H. Chen, "Kernel-Based Learning for Biomedical Relation Extraction," *J. Am. Soc. Information Science and Technology,* vol. 59, no. 5, pp. 756-769, 2008.
[18] T. Mitsumori, M. Murata, Y. Fukuda, K. Doi, and H. Doi, "Extracting Protein-Protein Interaction Information from Biomedical Text with SVM," *IEICE Trans. Information and Systems,* vol. E89D, no. 8, pp. 2464-2466, 2006.
[19] M. Yusuke, S. Kenji, S. Rune, M. Takuya, and T. Jun'ichi, "Evaluating Contributions of Natural Language Parsers to Protein-Protein Interaction Extraction," *Bioinformatics,* vol. 25, pp. 394-400, 2009.
[20] S. Novichkova, S. Egorov, and N. Daraselia, "MedScan, A Natural Language Processing Engine for MEDLINE Abstracts," *Bioinformatics,* vol. 19, no. 13, pp. 1699-1706, 2003.
[21] M. Ould Abdel Vetah, C. Nédellec, P. Bessières, F. Caropreso, A.-P. Manine, and S. Matwin, "Sentence Categorization in Genomics Bibliography: A Naive Bayes Approach," *Actes de la Journée Informatique et Transcriptome,* J.-F. Boulicaut and M. Gandrillon, eds., Mai 2003.
[22] J. Pustejovsky, J. Castaño, J. Zhang, M. Kotecki, and B. Cochran, "Robust Relational Parsing over Biomedical Literature: Extracting Inhibit Relations," *Proc. Pacific Symp. Biocomputing,* vol. 7, pp. 362-373, 2002.
[23] S. Ray and M. Craven, "Representing Sentence Structure in Hidden Markov Models for Information Extraction," *Proc. Int'l Joint Conf. Artificial Intelligence (IJCAI '01),* 2001.
[24] T.C. Rindflesch, L. Tanabe, J.N. Weinstein, and L. Hunter, "EDGAR: Extraction of Drugs, Genes, and Relations from the Biomedical Literature," *Proc. Pacific Symp. Biocomputing,* vol. 5, pp. 514-525, 2000.
[25] B. Rosario and M.A. Hearst, "Semantic Relations in Bioscience Text," *Proc. 42nd Ann. Meeting on Assoc. for Computational Linguistics,* vol. 430, 2004.
[26] P. Srinivasan and T. Rindflesch, "Exploring Text Mining from Medline," *Proc. Am. Medical Informatics Assoc. (AMIA) Symp.,* 2002.
[27] B.J. Stapley and G. Benoit, "Bibliometrics: Information Retrieval Visualization from Co-Occurrences of Gene Names in MEDLINE Abstracts," *Proc. Pacific Symp. Biocomputing,* vol. 5, pp. 526-537, 2000.
[28] J. Thomas, D. Milward, C. Ouzounis, S. Pulman, and M. Carroll, "Automatic Extraction of Protein Interations from Scientific Abstracts," *Proc. Pacific Symp. Biocomputing,* vol. 5, pp. 538-549, 2000.
[29] A. Yakushiji, Y. Tateisi, Y. Miyao, and J. Tsujii, "Event Extraction from Biomedical Papers Using a Full Parser," *Proc. Pacific Symp. Biocomputing,* vol. 6, pp. 408-419, 2001.

Oana Frunza received the BSc degree in computer science from Babes-Bolyai University, Romania, in 2004 and the MSc degree in computer science from the University of Ottawa, Canada, in 2006. Since 2004, when she joined the School of Information Technology and Engineering (SITE), University of Ottawa, she has been working as a research and teaching assistant. Currently, she is enrolled in the computer science PhD program in SITE. She has more than 10 publications in prestigious conferences and journals. She is also the author of the book "*Cognates, False Friends, and Partial Cognates*." Her research interests include artificial intelligence (AI), natural language processing (NLP), machine learning (ML), and text mining.

Diana Inkpen received the BEng from the Department of Computer Science, Technical University of Cluj-Napoca, Romania, in 1994, the MSc degree from the Technical University of Cluj-Napoca, Romania, in 1995, and the PhD degree from the Department of Computer Science, University of Toronto, in 2003. She worked as a researcher at the Institute for Computing Techniques in Cluj-Napoca, from 1996 to 1998. In 2003, after finishing the PhD degree, she joined the School of Information Technology and Engineering (SITE) at the University of Ottawa, as an assistant professor, and is currently an associate professor. She was an Erasmus Mundus visiting scholar, at the University of Wolverhampton, United Kingdom, during April-July 2009. She is a reviewer for several journals (*Computational Linguistics*, *Natural Language Engineering*, *Language Resources and Evaluation*, etc.) and a program committee member for many conferences (ACL, NAACL, EMNLP, RANLP, CICLing, TANL, AI, etc.). She published 4 book chapters, 11 journal papers, and 54 conference and workshop papers. She organized four international workshops. In 2005 and 2007, together with one of her PhD students, she obtained the best results at the Cross-Language Evaluation Forum, the Cross-Language Speech Retrieval track. Her research interests are in the areas of computational linguistics and artificial intelligence, more specifically: natural language understanding, natural language generation, lexical semantics, and information retrieval. She is member of the Association for Computational Linguistics (ACL).

Thomas Tran received the BSc (Four-Year Specialist) degree, double major in mathematics and computer Science from Brandon University in 1999, and the PhD degree in computer science from the University of Waterloo in 2004. He was the recipient of the Governor General's Gold Medal at the Convocation Ceremony. He worked as a postdoctoral researcher in the School of Computer Science, University of Waterloo from 2003 to 2004. He joined the School of Information Technology and Engineering (SITE), University of Ottawa as an assistant professor in 2004 and he is currently an associate professor. He is also a member of the Institute of Electrical and Electronics Engineers (IEEE), the Association for the Advancement of Artificial Intelligence (AAAI), and the Canadian Artificial Intelligence Association (CAIAC). He has had more than 40 publications in several refereed journals and conference proceedings, and he is the joint holder of the Outstanding Paper Award for a paper presented at IADIS E-Commerce-2008, the Best Paper Award Nominee at MCETECH-2009, and the Best Paper Award Nominee at UM-2003. His research interests include artificial intelligence (AI), electronic commerce, intelligent agents and multiagent systems, trust and reputation modeling, reinforcement learning, recommender systems, knowledge-based systems, architecture for mobile e-business, and applications of AI.