A Framework for Automated Corpus Generation for Semantic Sentiment Analysis

Amna Asmi and Tanko Ishaya, *Member, IAENG*

Abstract— **User-generated content is dominating a larger proportion of web resources, which has inspired research in other areas of web content analysis – such as sentiment analysis. The emergence of this user-generated web content provides analysis with both a vast corpus and a variety of potential applications. To achieve this analysis, a number of corpora have been developed and used for reference. While this has provided a mechanism for the analysis of sentiments at word level, further research is needed to advance analysis of sentiments to a semantic level. Since current approaches to the generation and annotation of existing corpora (such as SentiWordNet and Multi-Perspective Question Answering (MPQA)) is manually done, this can be time consuming and potentially generate conflicting understandings.**

This paper presents a proposed framework for automated generation of corpus based on analysis of opinions /sentiments and semantics in a user-generate free text. The framework has been developed based on an analysis of existing corpora – WordNet, SentiWordNet, Domain specific dictionaries, and Parts of Speech (POS) tagging mechanisms for syntactic and linguistic analysis.

Index Terms— **Automated corpus generation, semantic analysis, sentiment analysis, subjectivity**

I. Introduction

SENTIMENT analysis is an analysis of textual data, aiming to identify the positive and negative feelings, opinions, attitudes and emotions expressed within free text with respect to a target topic and the opinion holder (source). With the advent of web 2.0 and concept of user-generated content, a huge amount of free text is currently being generated on the web on a continuous basis and has made it possible for individuals outside the professional media to express their opinions on any topic that excites their interest. Commercial products (such as movie, music, books, travel destination, etc) reviews have so far played a central role in sentiment analysis [21]. In order to understand, analyse and utilise this text, a variety of analysis techniques are being use to derive useful meaning from these text. Due to the fundamental importance of words in any free text, initial text analysis usually starts with analysing how words are used, number of words in text, frequency of different words, co-occurrence of words, etc. Even most contemporary semantic analysis is based on words and keywords, which is why a variety of corpora (such as WordNet, SentiWordNet, and Multi-Perspective Question Answering (MPQA)) of words (positive/negative words, verbs/adverbs, etc.) and dictionaries (synonyms) are necessarily required for text analysis. Current corpora are all word based techniques, whereas the sentences generate meanings through words and their interaction with each other within and between the sentences. Since current approaches to the generation and annotation of existing corpora (semantic, syntactic, comparative linguistics) are manually done using multiple annotators. This can be time consuming and potentially generate inconsistencies and conflicting understandings. Furthermore, some of these corpora (especially SentiWordNet) are not large enough to incorporate all the worlds that might be encountered during the sentiment analysis. Thus the first step towards semantic analysis of sentiments is generating a complete corpus to be used for such analysis. This calls for the need of an alternative (automated) way to generate a corpus by defining set of rules. After evaluating some of the corpora, this paper presents a framework for automated generation of a corpus for semantic sentiment analysis of user-generate web content - such as forums.

The rest of the paper has been structured as follows: Section 2 presents an analysis of the related work in the areas of corpora generation and semantic and sentient analysis. This is followed by Section 3 describes the annotation scheme and the existing resources used in the generation (dependencies). Section 4 presents a proposed framework, which is followed by how this framework can be brought to existence. Section 5 involves the conclusion and the prospective future extensions of this work.

II. Related Work

The background to this paper is based on a review of existing corpora and techniques for sentiment analysis.

A. Review of Existing Corpora

Corpus related research is very diverse, especially in terms of linguistic based corpora. The efforts for creation of opinion mining and sentiment analysis have gained popularity within last three decades. The generation of reference corpus for training and testing of sentiment classifiers has been based on manual annotation of corpus which is really very expensive and time consuming process [8]. Generally, these corpora are domain specific and once they are created for one domain they cannot be shifted or used for other domains [27]. Intuitively, a robust text sentiment analyzer should be able to deal with a variety of domains; however, it is also clear that there is domain-specific information that cannot necessarily be handled by a

Manuscript submitted March 23, 2012;

A. Asmi is a PhD Candidate in Internet Computing with the University of Hull, Scarborough Campus, Filey Road, Scarborough YO11 3AZ, United Kingdom (phone: +44 (0) 1723 357235; e-mail: a.asmi@2008.hull.ac.uk).

T. Ishaya is with The University of Hull, Scarborough Campus, Filey Road, United Kingdom (email: t.ishaya@hull.ac.uk)

general system. MPQA is one of most renowned corpus for sentiment analysis [31]. It is developed and refined over time; although the annotations are based on word and phrase level not at sentence level [27]. Most of other researchers, who have tried to analyse sentiments have mostly annotated a set of sentences, or have generated their own lists/dictionaries/corpora to test their analysis techniques [18, 24, 34]. Another interesting way of corpus generation or to attest the corpus was undertaken by Pang and others in 2002, they have used movie reviews which were already rated by reviewers using star based rating. They are used for analysis, in order to test if the same rating is generated during analysis [24]. Later, Franky and Manurung (2008) have used same experiments as that of Pang (2002) but for Indonesian language [10]. They have used the machine translation tool to get a corpus in Indonesian language. The earliest efforts to get something for automated analysis or resource generation were from Sankoff and Cedegen in 1997, when they developed (varbrul) a program based on multivariate analysis technique for linguistic data [4]. Although, varbrul is still being reviewed and analysed by various researchers [12]. In 1991, a syntactic corpus of Middle English was presented. At that time, the technology was limited, however, in 1995 a tagger on the basis of lexicon and rules, was written by Bill [6]. Don Hindle developed 'fidditch' (an automated parser) for Modern English; later Mike Collins developed something better than it [4]. Automatic Mapping Among Lexico-Grammatical Annotation Models (AMALGAM) [17] is a research to generate a standard tagging scheme for English Language, by giving a mapper to map between main mapping schemes and parsing structures for grammar [4]. An international corpus of English (ICE) [4] is an ongoing process. It is a project having over 1 million English words. There was a try to generate a corpus for German, French and Italian in 1993/1996 [15]. Recent work for automated corpus generation uses tools like XML [5].

B. *Existing Techniques for Sentiment Analysis*

Hearst (1992) and Wiebe (1994) were the pioneers that proposed the idea of mining direction-based text. The direction-based text may include the opinions, sentiments, affects and biases [2]. Opinion mining in text analysis is almost similar text classification into positive and negative text, which can be done by different ways; on the basis of machine learning techniques as supervised/unsupervised techniques [24, 34] which include document level, sentence level or clause level techniques of sentiment mining [32, 33]. Whereas, some tried to find out the sentiment in multiple documents [25]. On the other hand, polarity, degree of polarity, features [16], subjectivity [13, 30], relationships identification [28], affect types [2, 22], mood classification [1, 21] and ordinal scales like giving ratings to reviews and products are used in some other researches [11]. However, these are all different view points towards the solution of same problem- sentiment analysis.

III. Annotation Process

The proposed framework explained in Section IV is implemented by generating a corpus. This corpus is a collection of cancer related forum data, which is taken from http://www.medhelp.org/forums/list. It is a limited data set, which is used to generate a corpus for the evaluation of proposed framework. This forum is in English, which has all the posts written as free text. Although, it has a set structure followed throughout the forum. The initial post is created by the author, who also gives it a subject. This post contains an issue/question also expressing the feelings about the experience. Subsequently, other members of forum reply and express their experiences and opinion about it. The members can also reply to any specific comment by mentioning the name of the member. In most of the cases, comments are more subjective text. The aim of this process is to summarise and annotate the opinion expressed in the text. In the next section, annotation scheme and methodology is discussed.

A. *Annotation methodology*

Annotation contains several steps, grabbing URL from which the threads will be extracted, author/commenter name, time of post, subject of post, text of post, number of comments and the information associated with comments. Text of the post is grabbed, it is broken into sentences. Each sentence is applied with Stanford Dependencies Parser and Penn Treebank Tagging. Then the sentences are broken on basis of their clauses (dependent/independent). Subject-Verb-Object triplet is extracted for each sentence. Some rules are specified based on POS (adjective/adverbs/verbs), negation, punctuations and conjunctions using SentiWordNet and WordNet. On the basis of specified rules, sentiments are extracted, polarity and its intensity is defined. Whereas, based on subject and object of the sentences and the topic/title of the forum/post, subjectivity is calculated. Figure 1 demonstrates the process of annotation, the interaction of the annotating tools explained in the next section and the resulting output.

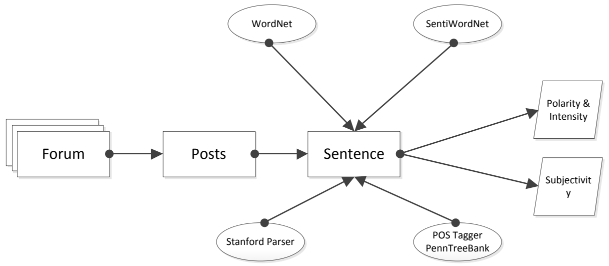

Fig. 1. Sentiment Analysis Process

B. *Tools for annotation*

WordNet

WordNet is an online lexical database for the English language developed at the Cognitive Science Laboratory of Princeton University [28]. In WordNet nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. It is designed with the consideration to be used by machines under program control [20] not only for humans. It is one of largest resources which provide intensive lexical coverage with semantic links among them. Although it does only base on single words, phrases and clauses are missing [7]. However, WordNet is ontologically organized with various relation links i.e. is-a) with each other. WordNet is planned to model the human glossary. Psycholinguistic findings have also been taken into account in its design phase [29]. WordNet keeps track of the context of situations in which words are being used, which provides help in defining semantically similar words as synonyms.

Primarily, it was used in our research for listing the synonyms of a word that is not found in the SentiWordNet. This helps a lot in giving a numeric value to the word, which will not be recognized in SentiWordNet. On the other hand, it helps in finding the relationships between words especially for subjectivity.

SentiWordNet

SentiWordNet was developed by taking the basic word set from WordNet, based on the assumption that same term can be used in different senses and they can have different opinions based on context and scenario [9, 22]. Esuli and Sebastiani (2006) have defined all the words based on the notions of “positivity”, “negativity”, and “neutrality”. Each sense is associated to three numerical scores Pos(s), Neg(s), and Obj(s) which indicate how positive, negative, and objective the term is. The sum of all three scores should add up to 1.0 [3, 9]. The proposed framework calculates the polarity of the thread. In this process the polarity of the words are taken from the SentiWordNet and are used for further calculation based on heuristic rules.

Stanford Parser

In natural language, the words have relations with each other within the sentences gives sense to the whole sentence. Stanford Parser is a probabilistic natural language parser; it works out the grammatical structure of sentences and uses knowledge of language gained from hand-parsed sentences to try to produce the most likely analysis of new sentences [14]. Stanford Parser is used in the proposed framework in order to get the dependency tree as it is one of the most recent and most used parser for dependency parsing.

PennTree Bank

The Penn Treebank Project was started by University of Pennsylvania. It annotates the naturally-occurring text for linguistic structure. It also annotates text with POS tags, which include 36 POS tags and 12 other tags (for punctuation and currency symbols) [19]. The proposed framework uses PennTree Bank to extract the subjects, objects and verbs from the sentence which are further used to generate triplet based on opinion holder, opinion and opinion topic from the sentence.

UMLS

UMLS is a system designed by US National Library of Medicine (NLM) to assist the computer systems to understand the meanings and the language of biomedicine terms [33]. UMLS Metathesaurus (an ontology) provides a rich lexicon which gives potential relationships such as (is-a and part-of) between concepts in UMLS [32]. UMLS is used in the proposed framework to find out the synonyms and definitions of medical terms and their relations with other medical terms.

IV. Framework

Figure 2 presents the functional architecture of the proposed framework to develop a corpus in the domain of medicine which is to provide the solution(s) to the problems in automatic understanding of semantics and sentiments in the posts of medical related forums. The storage data structure of the developed corpus enhances the efficiency and effectiveness in terms of information retrieval based on semantics and provides the evaluation of the retrieved information based on sentiments. The functionality of each module is discussed below.

A. *Repository*

The repository contains the knowledge used by the various modules introduced in the proposed framework. The repository includes WordNet [20] and SentiWordNet [9] dictionaries, Unified Medical Language System (UMLS) Metathesaurus [23] or other domain based resources, and rules for the sentence type identification, polarity identification, subjectivity identification and sentiment analysis.

B. *Data Pre-processor*

The proposed system takes the unstructured data from a medical forum (http://www.medhelp.org/forums/list) as input. The input is cleaned and filtered by the data pre-processor. The data pre-processor captures the thread structure of the forums and its comments, and arranges the captured information with the name of author, forum topic and data/time. It is followed by a process of spell check which is applied on the arranged information by the data pre-processor. The organized information is then split into a set of posts and passed onto the post pre-processor.

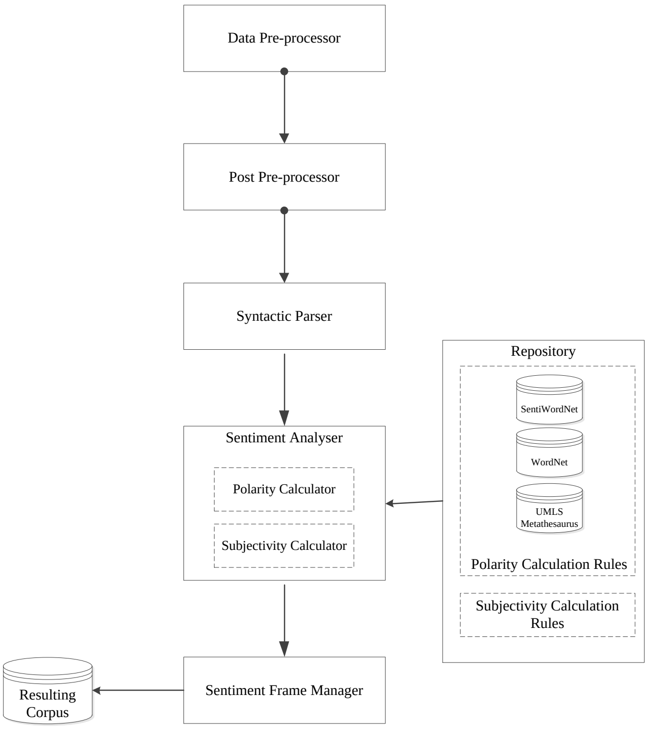

Fig. 2. Proposed Framework

C. *Post Pre-processor*

The post pre-processor obtains a list of posts in organised format. It splits the text into a set of sentences using Penn Tree Tagger. The post pre-processor then passes sentences to the syntactic parser iteratively as rest of the modules takes only single sentence to perform their functionality. Post pre-processor also keeps track the start and end of the post.

D. *Syntactic Parser (SP)*

The SP collects sentences iteratively and invokes a POS tagger. The POS tagger assigns a POS tag to each token in the sentence. The name entities and idioms involved in a sentence are also identified in syntactic parsing. The SP also identifies the dependencies/relationship within a sentence. The SP passes each sentence with all the identified information to the sentiment analyser after classifying the sentence as a question, an assertion, a comparison, a confirmation seeking or a confirmation providing by using the rule of sentence type identification. Figure 2 shows the graphical representation of the different steps performed by post pre-processor, SP and SA.

E. *Sentiment Analyser (SA)*

The SA considers as a heart of the proposed framework. The SA gathers sentences from the SP with all related information and extracts the sentiment oriented words from each sentence by using the relationship information of (dependencies within) the sentence. The Polarity Calculator (PC) and Subjectivity Calculator (SC) are the sub-modules of SA, which calculate the polarity and subjectivity of the each sentence respectively.

The PC calculates the polarity of the sentence and assigns a score. In order to calculate polarity, PC uses SentiWordNet to identify the positive and negative words through their values assigned by the SentiWordNet. In this process, PC collects the synonyms of the word if word is not found in SentiWordNet. The PC first uses WordNet to get the synonyms. However, the PC uses ULMS Metathesaurus if the synonyms are not found in the WordNet. The identified rules for polarity identification are also considered while polarity calculation.

The SC calculates the subjectivity of the sentence. The SC considers the POS and the relationships of the sentences for subjectivity calculation. The SC identifies all the sentences related to the topic of the post. The SC takes the nouns from the sentence and topic and associated information (synonyms, homonyms, meronyms, holonyms, hypernyms and hyponyms) from the WordNet for subjectivity identification.

The SA only takes polarities of those sentences which are related to the topic and marked by the SC, for the post polarity calculation. In order to calculate the post polarity, SA takes an aggregate of the polarities of the sentences related to the post. The SA generates the sentiment frame information for each sentence. A sentence may have more than one frame and a frame contains the type of sentence, subject, object/feature, sentiment oriented word(s), sentiment type (absolute or relative), sentiment strength (very weak, weak, average, strong or very strong), polarity of sentence, post index and sentence index. The SA forwarded the calculated polarity and subjectivity, and generated information for sentiment frames to the sentiment frame manager for the sentiment frame development and storage into a corpus.

F. *Sentiment Frame Manager*

The sentiment frame manager takes sentiment frame information generated by SA and stores all the information into a physical location (Resulted Corpus) as shown in Figure 2. The sentiment frame manager loaded all the frames into a tree structure from the Resulted Corpus at runtime memory on program load. It also keeps track the changes into loaded memory and appends those changes into Resulted Corpus after specific time span. The Resulted Corpus is automated and efficient in terms of storage and retrieval. It is stored into an XML file and on program load, complete corpus data is loaded into memory which reduces the retrieval time of the requested data from the corpus.

V. Conclusion & Future Work

While current approaches for sentiment analysis has provided a mechanism for the analysis of sentiments at word level, further research is needed to advance analysis of sentiments to a semantic level. This paper presents a proposed framework for automated generation of corpus based on analysis of opinions /sentiments and semantics in a user-generate free text. The framework has been developed through the analysis of existing corpora – WordNet, SentiWordNet, Domain specific dictionaries, and POS tagging mechanisms for syntactic and linguistic analysis. In general, this framework aims to give computer program a skill to recognize and understand the emotion hidden in a sentence, and correlate it with other sentences of the paragraph. It not only will sense the sentiment within the text but also will evaluate the strength. It will analyze the sentence lexically and semantically. The rules for sentence clauses, negation and punctuation are considered especially. For further work the analysis for sentiments and subjectivity is required to make it general purpose.

The developed framework is currently being evaluated using a medical-based forums incorporating UMLS corpus. The generic nature of the framework will also be proved in the future work as the proposed framework is currently working and evaluated with the medical forum.

References

[1] Abbasi, A. (2007) Affect Intensity Analysis of Dark Web Forums. In: Intelligence and Security Informatics, 2007 IEEE. 282-288.

[2] Abbasi, A., Chen, H. and Salem, A. (2008) Sentiment analysis in multiple languages: Feature selection for opinion classification in Web forums. ACM Trans. Inf. Syst., Vol. 26, 3, pp. 1-34.

[3] Baccianella, S., Esuli, A. and Sebastiani, F. (2010) SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In: The Seventh conference on International Language Resources and Evaluation LREC'10. Valletta, Malta. European Language Resources Association ELRA (May 2010)

[4] Bhattacharyya, D., et al. (2009a) Refine Crude Corpus for Opinion Mining. In: The 2009 First International Conference on Computational Intelligence, Communication Systems and Networks. Washington, DC, USA. IEEE Computer Society, 17-22.

[5] Bhattacharyya, D., et al. (2009b) An Approach of XML-ifying the Crude Corpus in the Field of Opinion Mining. Distributed Computing, Vol. 2, 3, pp. 13-24.

[6] Breck, E., Choi, Y. and Cardie, C. (2007) Identifying expressions of opinion in context. In: The 20th international joint conference on Artifical intelligence. Hyderabad, India. Morgan Kaufmann Publishers Inc, 2683-2688.

[7] Chow, I. C. and Webster, J. J. (2009) Supervised Clustering of the WordNet Verb Hierarchy for Systemic Functional Process Type Identification.

[8] Devitt, A. and Ahmad, K. (2008) Sentiment Analysis and the Use of Extrinsic Datasets in Evaluation. In: The International Conference on Language Resources and Evaluation, LREC 2008. Marrakech, Morocco. European Language Resources Association, 1063-1066.

[9] Esuli, A. and Sebastiani, F. (2006) SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining. In: 5th Conference on Language Resources and Evaluation. Genova, IT.

[10] Franky and Manurung, R. (2008) Machine learning-based sentiment analysis of automatic indonesian translations of english movie reviews. In: The International Conference on Advanced Computational Intelligence and Its Applications 2008 (ICACIA 2008). Depok, Indonesia.

[11] Ghose, A., Ipeirotis, P. and Sundararajan, A. (2007) Opinion Mining using Econometrics: A Case Study on Reputation Systems. In: Proceedings of ACL-07, the 45th Annual Meeting of the Association of Computational Linguistics. Association for Computational Linguistics, 416-423.

[12] Gorman, K. (2009) On VARBRUL - Or, The Spirit of `74. Philadelphia, PA, USA, Institute for Research in Cognitive Science, University of Pennsylvania.

[13] Hatzivassiloglou, V. and Wiebe, J. M. (2000) Effects of adjective orientation and gradability on sentence subjectivity. In: Proceedings of the 18th conference on Computational linguistics - Volume 1. Saarbr\&\#252;cken, Germany. Association for Computational Linguistics,

[14] Ims_Corpus (1993) Textcorpora und Erschliessungswerkzeuge ('textual corpora and tools for their exploration') [WWW]. Available from: http://www.ims.uni-stuttgart.de/projekte/tc/ [Accessed March 18, 2012].

[15] Joshi, M. and Penstein-Ros, C. (2009) Generalizing dependency features for opinion mining. In: Proceedings of the ACL-IJCNLP 2009 Conference Short Papers. Suntec, Singapore. Association for Computational Linguistics,

[16] Leeds_University (n.d) Automatic Mapping Among Lexico-Grammatical Annotation Models (AMALGAM) [WWW]. Available from: http://www.scs.leeds.ac.uk/ccalas/amalgam/amalgover.html [Accessed March 18, 2012].

[17] Li, N. and Wu, D. D. (2010) Using text mining and sentiment analysis for online forums hotspot detection and forecast. Decision Support Systems, Vol. 48, 2, pp. 354-368.

[18] Miller, G. A. (1995) WordNet: A Lexical Database for English. Communications of The ACM Vol. 38, 11, pp. 39-41.

[19] Mishne, G. (2006) AutoTag: a collaborative approach to automated tag assignment for weblog posts. In: Proceedings of the 15th international conference on World Wide Web. Edinburgh, Scotland. ACM,

[20] Neviarouskaya, A., Prendinger, H. and Ishizuka, M. (2009) SentiFul: Generating a reliable lexicon for sentiment analysis. In: Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on. 1-6.

[21] Pang, B., Lee, L. and Vaithyanathan, S. (2002) Thumbs up?: sentiment classification using machine learning techniques. In: The ACL-02 conference on Empirical methods in natural language processing Philadelphia, PA, USA. Association for Computational Linguistics, 79-86.

[22] Riloff, E., Wiebe, J. and Phillips, W. (2005) Exploiting subjectivity classification to improve information extraction. In: Proceedings of the 20th national conference on Artificial intelligence - Volume 3. Pittsburgh, Pennsylvania. AAAI Press,

[23] Sarmento, L., et al. (2009) Automatic creation of a reference corpus for political opinion mining in user-generated content. In: The 1st international CIKM workshop on Topic-sentiment analysis for mass opinion. Hong Kong, China. ACM, 29-36.

[24] Silva, M. J., et al. (2009) The Design of OPTIMISM, an Opinion Mining System for Portuguese Politics. In: New Trends in Artificial Intelligence: Proceedings of EPIA 2009 - Fourteenth Portuguese Conference on Artificial Intelligence. Aveiro, Portugal. Universidade de Aveiro, 565-576.

[25] Somasundaran, S., et al. (2006) Manual annotation of opinion categories in meetings. In: Proceedings of the Workshop on Frontiers in Linguistically Annotated Corpora 2006. Sydney, Australia. Association for Computational Linguistics,

[26] Suchanek, F. M., Kasneci, G. and Weikum, G. (2007) YAGO: A Core of Semantic Knowledge Unifying WordNet and Wikipedia. In: The 16th international conference on World Wide Web. Banff, Alberta, Canada. ACM,

[27] Wiebe, J. and Riloff, E. (2011) Finding Mutual Benefit between Subjectivity Analysis and Information Extraction. Affective Computing, IEEE Transactions on, Vol. PP, 99, pp. 1-1.

[28] Wiebe, J., Wilson, T. and Cardie, C. (2005) Annotating expressions of opinions and emotions in language. Language Resources and Evaluation, Vol. 39, 2-3, pp. 165-210.

[29] Wiebe, J. M., Bruce, R. F. and O'hara, T. P. (1999) Development and use of a gold-standard data set for subjectivity classifications. In: Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics. College Park, Maryland. Association for Computational Linguistics,

[30] Wilson, T., Wiebe, J. and Hwa, R. (2004) Just how mad are you? finding strong and weak opinion clauses. In: Proceedings of the 19th national conference on Artifical intelligence. San Jose, California. AAAI Press,

[31] Yu, H. and Hatzivassiloglou, V. (2003) Towards answering opinion questions: separating facts from opinions and identifying the polarity of opinion sentences. In: The 2003 Conference on Empirical Methods in Natural Language Processing (EMNLP 2003) Sapporo, Japan. Association for Computational Linguistics, 129-136.

[32] Berlanga, R., et al. (2008) Medical Data Integration and the Semantic Annotation of Medical Protocols. In: The 2008 21st IEEE International Symposium on Computer-Based Medical Systems. Jyväskylä, Finland. IEEE Computer Society, 644-649.

[33] Nlm.Nih.Gov (2012) Unified Medical Language System® (UMLS®) [WWW]. Available from: http://www.nlm.nih.gov/pubs/factsheets/umls.html [Accessed April 14, 2012].