Looking for Subjectivity in Medical Discharge Summaries
The Obesity NLP i2b2 Challenge (2008)

Michael Roylance and Nicholas Waltner

Abstract

This paper focuses on the task of text classification ("TC") within the medical record domain. Specifically, we examined 1,237 discharge summaries of patients diagnosed with obesity and 15 associated co-morbidities. The description and results of the initial competition are described in the Uzuner paper (2008) summarizing the i2b2 Obesity Challenge. We decided to take the route of determining whether sentimental metrics could be applied to machine learning features for improved classification accuracy. In addition to the usage of standard computational linguistic tools, such as the Stanford Parser and Mallet, we also employed the Genia Tagger trained specifically for parsing bio-medical texts. The original competition consisted of four different classifications for each disease: Present, Absent, Questionable and Unmentioned. In order to narrow the scope of our project, we limited the classes to simply Present and Absent. As pointed out in the Uzuner paper, the top systems were designed with extensive knowledge of the medical domain. However, as non-medical professionals with a curiosity to see how we could compare with standard text mining procedures, we applied a wide range of sentiment features. Our aggregate best F-scores across all 16 diseases for intuitive and textual were 56.4% and 61.9%, which represent 4.6% and 5.5% improvement versus our two-class baseline. However, our techniques did help improve accuracy in co-morbidity accuracy cases such as Depression, Diabetes, and Gout where there was a relatively high level of disagreement between *textual* and *intuitive* class labeling. Although the addition of sentiment features was not found to have improved upon the classification accuracy of the obesity corpus, we did detect a very strong relationship between the co-morbidities, which may lead to a different approach to the classification problem in the future.

1 Introduction

In this paper we examined whether the methods learned in Linguistics 575 could be effectively applied to the domain of medical records. At a minimum, the domain of medical records presents a number of special challenges in terms of natural language processing techniques including:

- Intensive use of highly specialized medical terms[1].
- Common use of abbreviations.
- Challenging issues in named entity recognition (NER).
- Non-standard grammatical expressions.
- Inconsistent free text fields, which are governed by headers that can differ in location and content.

Although the teams in the competition spent significant resources in dealing with these basic NLP tasks within the text pre-processing leg of their systems, we largely opted not to deal with these challenges, with the exception of a few word count features where some of the part of speech tagged words were easily identifiable as noise. As such, we decided to turn our focus to the potential sentential aspects of the corpus.

[1] Many of the teams participating in the Obesity Challenge relied on the UMLS or Unified Medical Language System, which is a compilation of the many controlled vocabularies in the biomedical sciences and provides mappings between various medical terms, in order to expand, contract or resolve anaphoric references to medical conditions and diseases

1.1 Obesity Challenge

The underlying focus of our project was that of the i2b2 Obesity Challenge, which was sub-termed as *"Who's obese and what co-morbidities do they (definitely/likely) have?"*. This somewhat pun-in-check title speaks to the statistical fact that many patients diagnosed with obesity also suffer from a number of other diseases, which in the medical profession are termed co-morbidities. An initial investigation of the data set revealed that, in fact, each patient had been diagnosed with 4.37 co-morbidities on average, which potentially spoke to a clustering of diseases and/or disease diagnosis confusion.

The obesity challenge itself consisted of two annotation data sets (textual and intuitive) along with classification of whether 16 diseases were Present, Absent, Questionable or Unmentioned. Given the low frequency of the Absent and Questionable classes, the competition was based on micro- and macro-average precision, recall and F1-scores versus simple accuracy measures in most test classification tasks. The data set itself consisted of the discharge summaries from Partners Healthcare. The textual annotations were made on the basis of the presence of certain textual clues, whereas the intuitive annotations were made on the basis of careful human review of the records. Below in Tables 1. and 2., we provide a summarization of the two data sets:

Table 1: Distribution of Textual Judgements into Training and Test Sets

	Present		**Absent**		**Questionable**		**Unmentioned**		**Total**	
Diseases	**Training**	**Test**	**Training**	**Test**	**Training**	**Test**	**Training**	**Test**	**Training**	**Test**
Asthma	93	68	3	2	2	2	630	432	728	504
CAD	399	277	23	22	7	2	292	196	721	497
CHF	310	205	11	11	0	0	399	280	720	496
Depression	104	72	0	0	0	0	624	434	728	506
Diabetes	485	338	15	12	7	3	219	150	726	503
GERD	118	69	1	1	5	1	599	433	723	504
Gallstones	109	87	4	2	1	0	615	418	729	507
Gout	90	52	0	0	4	0	634	453	728	505
Hypercholesterolemia	304	213	13	6	1	4	408	279	726	502
Hypertension	537	374	12	6	0	0	180	121	729	501
Hypertriglyceridemia	18	10	0	0	0	0	711	497	729	507
OA	115	86	0	0	0	0	613	416	728	502
OSA	105	69	1	0	8	2	614	432	728	503
Obesity	298	198	4	3	4	3	424	289	730	493
PVD	102	64	0	0	0	0	627	443	729	507
Venous.Insufficiency	21	10	0	0	0	0	707	497	728	507
Total	3,208	2,192	87	65	39	17	8,296	5,770	11,630	8,044

Notes: CAD = coronary artery disease; CHF = congestive heart failure; DM = diabetes mellitus; GERD = gastroesophageal reflux disease; HTN = hypertension; OSA = obstructive sleep apnea; OA = osteo arthritis; PVD = peripheral vascular disease.

Table 2: Distribution of Intuitive Judgements into Training and Test Sets

	Present		Absent		Questionable		Unmentioned		Total	
Diseases	**Training**	**Test**	**Training**	**Test**	**Training**	**Test**	**Training**	**Test**	**Training**	**Test**
Asthma	86	68	596	403	0	0	0	0	682	471
CAD	391	272	265	185	5	1	0	0	661	458
CHF	308	205	318	229	1	4	0	0	627	438
Depression	142	105	555	372	0	0	0	0	697	477
Diabetes	473	333	205	146	5	0	0	0	683	479
GERD	144	93	447	331	1	2	0	0	592	426
Gallstones	101	80	609	411	0	0	0	0	710	491
Gout	94	61	616	439	2	0	0	0	712	500
Hypercholesterolemia	315	242	287	189	1	0	0	0	603	431
Hypertension	511	358	127	88	0	0	0	0	638	446
Hypertriglyceridemia	37	25	665	461	0	0	0	0	702	486
OA	117	91	554	367	1	4	0	0	672	462
OSA	99	66	606	427	8	2	0	0	713	495
Obesity	285	192	379	255	1	0	0	0	665	447
PVD	110	65	556	399	1	1	0	0	667	465
Venous.Insufficiency	54	29	577	398	0	0	0	0	631	427
Total	3,267	2,285	7,362	5,100	26	14	0	0	10,655	7,399

Notes: CAD = coronary artery disease; CHF = congestive heart failure; DM = diabetes mellitus; GERD = gastroesophageal reflux disease; HTN = hypertension; OSA = obstructive sleep apnea; OA = osteo arthritis; PVD = peripheral vascular disease.

1.2 Motivation

Given Michael's familiarity with the bio-medical realm and Nick's enthusiasm for text classification, the Obesity data set seemed like a great challenge, in order to evaluate with various measures of sentiment and subjectivity, which could aid in the task of improving the classification accuracies of the 16 diseases across both of annotation data sets, especially in the case of the intuitive one.

Another motivation behind the paper was to determine whether the language used to describe and communicate patient information could be categorically scanned for emotional content and whether such emotional content could provide a hidden pattern for useful features for disease classification. For example, the overuse of negative words to describe a patient's condition might also exhibit a co-morbidity of depression. This model of sentimental descriptions might also be a useful data point to healthcare providers and insurance companies.

As we mentioned above, the challenges of the bio-medical textual domain are quite formidable and we certainly were challenged by them in this project. In the end we came to the conclusion that sentential aspects in the medical corpus did not improve classification results owing to:

- Above all, a general lack of emotive terms within the patient discharge summaries.
- A high level of Part-of-Speech tagging confusion, which was improved by the use of the Genia Tagger, but still was nonetheless an issue.
- A high degree of medical domain specific terminology, which we did not have time to consolidate via regexes or UMLS contractions/expansions.
- Finally, an insufficiently sized training corpus in order to adequately model the 48 classifications (16 diseases by 4 categories).

The highest performing systems in the competition, as described in the Uzuner paper, were ones that utilized rule based heuristics generated from expert knowledge of the medical domain. Our system, in contrast, was constructed largely with the provided data set with our engineered features, the Stanford and Genia parser outputs and the Subjective Lexicon.

Although we were disappointed by our results, this could be considered an encouraging argument for both the capabilities of the human mind as it is compared to the growing functionality of machine learning and artificial intelligence, as well as a validation of the effectiveness of current medical profession training. Our system was not able to outperform the best heuristic based systems built for the competition, which may point to the tangible value that the data generated from experience and training brings to the accuracy of disease classification in patients.

Table 3: Competition Results - Textual Annotations

	Macro Averaged			**Micro Averaged**		
Team	**Precision**	**Recall**	**F-Score**	**Precision**	**Recall**	**F-Score**
Yang et al.	0.8482	0.7737	0.8052	0.9723	0.9723	0.9723
Solt et al.	0.8318	0.7776	0.8000	0.9756	0.9756	0.9756
Ware et al.	0.8314	0.7542	0.7821	0.9718	0.9718	0.9718
Childs et al.	0.8169	0.7454	0.7762	0.9773	0.9773	0.9773
Mishra et al.	0.7485	0.8050	0.7718	0.9704	0.9704	0.9704
Szarvas et al.	0.7644	0.7600	0.7622	0.9729	0.9729	0.9729
Savova et al.	0.7701	0.7147	0.7377	0.9668	0.9668	0.9668
Patrick et al.	0.7971	0.6219	0.6737	0.9693	0.9693	0.9693
Jazayeri et al.	0.7849	0.5779	0.6205	0.9514	0.9514	0.9514
DeShazo et al.	0.8552	0.6240	0.6140	0.9639	0.9639	0.9639

Table 4: Competition Results - Intuitive Annotations

	Macro Averaged			**Micro Averaged**		
Team	**Precision**	**Recall**	**F-Score**	**Precision**	**Recall**	**F-Score**
Solt et al.	0.7485	0.6571	0.6745	0.959	0.959	0.959
Szarvas et al.	0.6999	0.6588	0.6727	0.9642	0.9642	0.9642
Childs et al.	0.7061	0.654	0.6696	0.9582	0.9582	0.9582
Ware et al.	0.641	0.6399	0.6404	0.9654	0.9654	0.9654
Ambert et al.	0.6383	0.6307	0.6344	0.9558	0.9558	0.9558
Meystre	0.6304	0.6387	0.6343	0.9566	0.9566	0.9566
Yang et al.	0.6383	0.6294	0.6336	0.9572	0.9572	0.9572
†DeShazo et al.	0.9722	0.6216	0.6292	0.9524	0.9523	0.9524
Matthews	0.6325	0.6256	0.6288	0.9509	0.9509	0.9509
Jazayeri et al.	0.632	0.6257	0.6287	0.9508	0.9508	0.9508

2 Related Work

We reviewed the medical text classification papers mentioned in our bibliography; however, we found that a careful review of Uzuner along with a number of the papers of the top-ranked teams in the challenge provided very deep insights into the different teams methodologies and hence possible tasks for improvement. Specifically, Yang et al discuss their use of the Genia Tagger, while the Mayo Clinic team discussed the use of multiple classification techniques depending on the disease training from decision trees, maximum entropy and support vector machines. While we did not review all of the submission papers, we did not notice a discussion of sentiment related features in any of the papers.

Further, as we found in our literature review, there has been some work done in the classification of patient sentiment within suicide notes[2] by medical professionals; however, the language utilized in such notes is quite different than the type used in discharge summaries of our data set, i.e. consisting of more standard vernacular sentences.

3 Methodology

First, we broke down each record into sentences and then used the Stanford part of speech tagger to identify each word's lexical class. We stored this information in two tables in MySql with one for sentences containing the POS tagged sentence with a manually generated ID for the global sentence number, and then a more granular part of speech table, which contained the sentence ID, the word and the part of speech. We also added in an additional column, termed the part of speech convert, which we would later use to join the subjective words and parts of speech together. We also added indexes for all of the columns for faster searching capabilities through MySql, which proved extremely useful in delving into the finer grained aspects of the data set.

As per Professor Levow's suggestion and as indicated in Yang et al, we also employed the Genia Tagger in addition to the Stanford one. We followed the same process with this tagged corpus and loaded the data into two other tables in the database. This set also gave us easily identifiable medical named entities.

Next, we used a popular sentiment analysis lexicon, Subjective Lexicon, to help identify polarity in the text. This lexicon was used in conjunction with the previously loaded tables to identify words and their parts of speech which could have sentimental polarity.

We used Python to pull the data down from MySql with specific queries for the record classification and word counts. We used Mallet to build our trainable instances for classification. We created a custom Python framework to easily pipeline the addition of new feature vectors, the calling of Mallet with the instanced features, the extraction of results from Mallet, and the upload of the results to a common MySql table for quick comparison.

This framework can be found on Mike's Github site, although the Genia, Mallet, and Stanford applications are not included in the repository: `https://github.com/roylanceMichael/Partners_Obesity_Subjective_Feature_Learning`

Below is a list of the features we built and tested in our framework:

[2] http://www1.cs.columbia.edu/ julia/papers/suicide.pdf

Table 5: Classification Feature

Feature	Description
Word Count	A standard word count that grabs each word and its count and builds a vector.
Word Count Filtered	The same word count algorithm, but ignoring the "X" part of speech as identified by the Stanford part of speech tagger.
Genia Nouns	Base words that are are identified by Genia as nouns followed by the number of times that word/noun appeared in the document. This was also accompanied by the standard Word Count.
Genia Verbs	Base words that are identified by Genia as verbs followed by the number of times that word/verb appeared in the document. This was also accompanied by the standard Word Count.
Genia Pronouns	Base words that are identified by Genia as pronouns followed by the number of times that word/pronoun appeared in the document. This was also accompanied by the standard Word Count.
Genia Adjectives	Base words that are identified by Genia as adjectives followed by the number of times that word/adjective appeared in the document. This was also accompanied by the standard Word Count.
Genia Adverbs	Base words that are identified by Genia as adverbs followed by the number of times that word/adverb appeared in the document. This was also accompanied by the standard Word Count.
Genia Adverbs/Adjectives	Base words that are identified by Genia as adverbs/adjectives followed by the number of times that word/adverb/adjective appeared as in the document. This was also accompanied by the standard Word Count.
Genia Part of Speech	Parts of speech identified by Genia and the number of times they appeared in the document identified.
Stanford Nouns	Words that are are identified by Stanford as nouns followed by the number of times that word/noun appeared in the document. This was also accompanied by the standard Word Count.
Stanford Verbs	Words that are identified by Stanford as verbs followed by the number of times that word/verb appeared in the document. This was also accompanied by the standard Word Count.
Stanford Pronouns	Words that are identified by Stanford as pronouns followed by the number of times that word/pronoun appeared in the document. This was also accompanied by the standard Word Count.
Stanford Adjectives	Words that are identified by Stanford as adjectives followed by the number of times that word/adjective appeared in the document. This was also accompanied by the standard Word Count.
Stanford Adverbs	Words that are identified by Stanford as adverbs followed by the number of times that word/adverb appeared in the document. This was also accompanied by the standard Word Count.
Stanford Adverbs/Adjectives	Words that are identified by Stanford as adverbs/adjectives followed by the number of times that word/adverb/adjective appeared in the document. This was also accompanied by the standard Word Count.
Stanford Part of Speech	Parts of speech identified by Stanford and the number of times they appeared in the document identified.
Genia Polarity	Base words and their polarities identified by Genia and Subjective Lexicon as having a positive, negative, neutral or both polarities in the document.
Genia Polarity Type	Base words and their polarities and types identified by Genia and Subjective Lexicon as having positive, negative, neutral or both and whether the polarity is considered weak or strong in the document.
Stanford Polarity	Words and their polarities identified by Stanford and Subjective Lexicon as having a positive, negative, neutral or both polarities in the document.
Stanford Polarity Type	Base words and their polarities and types identified by Stanford and Subjective Lexicon as having positive, negative, neutral or both and whether the polarity is considered weak or strong in the document.
Genia Named Entity	Named entities identified by Genia and their counts for each document.
Genia Chunk	High level phrasal chunks (IE Noun Phrase, Verb Phrase etc) and the counts in the document.

4 Results and Discussion

The highest performing features were the Genia POS Noun feature, for both intuitive and textual. This feature took a list of the base words also identified by nouns and printed out its count in addition to the word count feature. It is also interesting to note that the Genia Tagger identified 618,342 words that were nouns, as compared with 370,390 nouns in the Stanford parser. However, the lopsided tagging did not seem to diminish the other scores for Genia as it compared with Stanford well.

The standard word count feature has been known to be a strong performer and our tests support that. The standard word count had some of our highest results, with 53.9 and 58.7 percent. However, once we filtered out the noise tags (X), the word count received dramatically lower scores. Those tags that we thought were noise turned out to produce better results when included.

Table 6: Classification Results with Present / Not Present

Category	Sub-Task	Intuitive F-Score / Accuracy	Textual F-Score / Accuracy
Base Line	Word Count	53.9/**85.8**	58.7/**88.7**
	Word Count Filtered	43.5/81.6	45.9/83.8
Stanford POS	Pronouns	53.5/85.6	58.3/88.5
	Nouns	53.5/85.2	56.7/87.6
	Verbs	50.6/84.9	54.8/87.7
	Adjectives	52.2/85.0	56.2/87.8
	Adverbs	53.4/85.3	57.2/88.2
	Adjectives and Adverbs	51.0/84.8	55.4/87.5
	All Tags	54.1/**85.8**	57.8/88.4
Genia POS	Pronouns	53.9/85.6	58.5/88.6
	Nouns	**56.4/86.2**	**61.9/89.1**
	Verbs	52.5/85.3	56.3/87.9
	Adjectives	53.9/85.2	57.8/88.1
	Adverbs	52.6/85.0	57.4/88.0
	Adjectives and Adverbs	52.5/84.8	56.5/87.7
	All Tags	44.3/81.8	45.4/83.8
Sentiment	Stanford Polarity	43.6/81.6	45.7/83.7
	Stanford Polarity Type	44.0/81.7	45.4/83.7
	Genia Polarity	43.5/81.6	45.8/83.7
	Genia Polarity Type	43.7/81.7	46.3/83.8
Genia Features	Named Entity	54.2/85.6	58.0/88.4
	Chunk	54.0/**85.7**	58.4/88.5

The sentiment scores indicate that our inclusion of them does not better help classify the instances. Compared to the results from individual part of speech tagged features for both Stanford and Genia, polarity did not make much difference.

Several of the features did report high F-Scores on certain diseases:

Table 7: Highest Performing Features by Disease

Source	Feature	Disease	F-Score
Textual	Genia Verbs	OA	91.6
Textual	Genia Nouns	OA	91.5
Textual	Genia Nouns	Hypertension	91.4
Textual	Stanford Pronouns	CHF	90.0
Textual	Genia Chunk	Diabetes	88.2
Textual	Genia Nouns	CAD	87.9

However, the high scores were offset by the lowest performing scores:

Table 8: Lowest Performing Features by Disease

Source	Feature	Disease	F-Score
Textual	Genia All Tags	GERD	13.3
Textual	Stanford Polarity Type	GERD	15.4
Textual	Stanford Polarity Type	Gallstones	15.4
Textual	Stanford Polarity Type	Depression	16.5
Textual	Stanford Polarity Type	Venous Insufficiency	16.7

The low scores might indicate a lack of data. However, there are diseases that perform well despite having few instances.

Table 9: Disease Training Instance Count

Disease Name	Number of records classified
Hypertension	1048
Diabetes	958
CAD	790
Hypercholesterolemia	619
CHF	618
Obesity	583
GERD	262
Depression	246
OA	232
PVD	212
Gallstones	210
OSA	204
Gout	184
Asthma	179
Venous Insufficiency	75
Hypertriglyceridemia	55

Certain diseases with high average F-score did not necessarily have the highest amount of training instances. For instance, OA was consistently high, ranging from 89.2 to 91.6, depending on the feature, despite only having 232 instances to train. Meanwhile GERD ranged from 13.3 to 49.5, depending on the feature. The variation of F-Score in certain diseases is a concern.

Table 10: Disease FMeasure Scores

Disease Name	Average F-Score	F-Score Standard Deviation
OA	90.7	0.74
Hypertension	88.5	2.1
CHF	87.3	4.0
CAD	83.8	4.9
Diabetes	74.1	2.3
Obesity	66.9	19.4
Hypercholesterolemia	60.1	17.7
OSA	45.8	15.8
Asthma	43.4	13.8
PVD	41.8	13.9
Gout	37.0	13.0
GERD	32.1	14.6
Depression	27.8	11.4
Gallstones	19.9	6.7
Hypertriglyceridemia	16.7	0
VenousInsufficiency	16.7	0

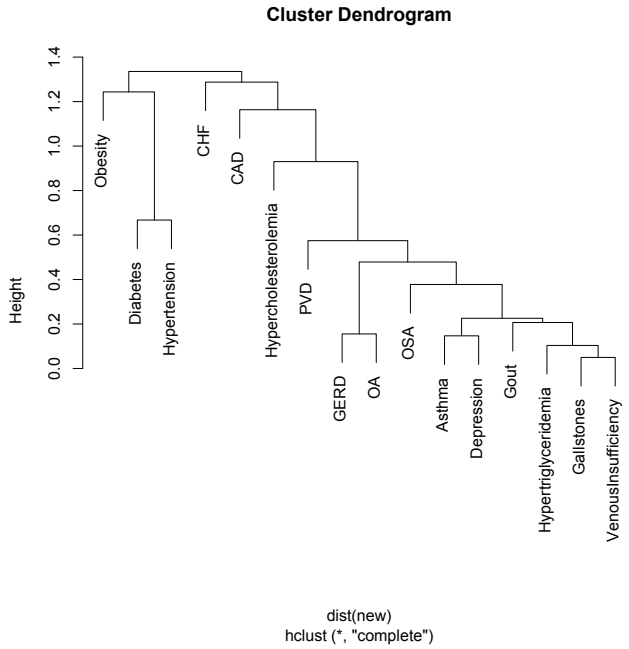

Figure 1: Cluster Dendrogram

Given our disappointing results, we examined the co-occurence of co-morbidities around the entire dataset. Above in Figure 1, we provide a the results of a hierarchical clusters analysis of the 16 diseases based upon on principal component extraction of the first five factors from the data set. As one can see, the sixteen diseases split into two major clusters with one being very closely related to obesity, i.e. Diabetes and Hypertension, while the other 13 diseases seem to relate to different levels of heart-related dysfunction.

5 Conclusion and Future Work

Our goal for this project was to determine whether we could train a system to learn the patterns of disease diagnosis given the information provided in the patient discharge records that competed favorably with others that expert medical professionals had created. As noted, the best approaches from the competition were systems built with heuristics working directly with medical professionals. What that means in relation to our work was that the data set used to calculate these results was much larger for these individuals, as the many years of experience helped filter out the noise for effective rule based parsing. Our system, in contrast, took standard approaches of breaking the words and sentences down to their widely recognizable English compositions and attempted to identify patterns from simple constructions of those compositions.

Future work would include identifying the variation in certain disease / feature set combinations and altering the system to successfully reduce it in certain diseases without increasing it in others.

6 Acknowledgments

We appreciate the use of various open source technologies, such as MySql, Python, SKLearn, Numpy, Mallet and Latex among others.
Michael handled the data extraction, MySql population and feature engineering.
Nick was responsible for the in class presentations and for the rough draft of the final paper.

References

[1] Savova G, Clark C, Zheng J, et al. *The Mayo/MITRE system for discovery of obesity and its comorbidities*. Proceedings of the i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data 2008.

[2] Szarvas G. *Automatic obesity-related morbidity.* http://www.inf.u-szeged.hu/rgai/?lang=en&page=obesity 2009. Accessed: Apr 2.

[3] Uzuner, O. *Recognizing Obesity and Comorbidities in Sparse Data*. J Am Med Inform Assoc. 2009 Jul-Aug; 16(4): 561–570. doi: 10.1197/jamia.M3115.

[4] Yang H, Spasic I, Keane JA, Nenadic G. *A text mining approach to the prediction of a disease status from clinical discharge summaries*. J Am Med Inform Assoc 2009;16:596-600.

[5] Lehnert W, Soderl S, Aronow D, Feng F and Shmueli A. *Inductive Text Classification for Medical Applications.* Journal for Experimental and Theoretical Artificial Intelligence 7(1), pp. 271-302, 1995.

[6] Claster, W., Shanmuganathan, S., and Ghotbi, N. (2007). *A Hybrid Approach to Data Mining Radiological Medical Records*. In G. Vidmar, and J. Stare (Eds.), Proceedings of the INTERNATIONAL Conference Applied Statistics. Ljubljana: Statistical Society of Slovenia.

[7] Theresa Wilson, Janyce Wiebe, and Paul Hoffmann (2005) *Recognizing Phrase Level Polarity in Phrase-Level Sentiment Analysis, Proc. of HLT-EMNLP-2005.*

[8] Richard Socher, John Bauer, Christopher D. Manning and Andrew Y. Ng. (2013). *Parsing with Compositional Vector Grammars, Proceedings of ACL 2013*

[9] Ashish Sureka, Vikram Goyal, Denzil Correa, and Anirban Mondal (2009) *Polarity Classification of Subjective Words Using Common-Sense Knowledge(2009).*

[10] Riloff, Wiebe and Wilson. (2003) *Learning Subjective Nouns using Extraction Pattern Bootstrapping.*