WME 3.0:
An Enhanced and Validated Lexicon of Medical Concepts

Anupam Mondal, Dipankar Das, Erik Cambria, and Sivaji Bandyopadhyay

Sentic Team

http://sentic.net

SenticNet for Business

IEEE IS ACSA

Department of
IEEE Intelligent Systems on
Affective Computing and Sentiment Analysis

http://sentic.net/acsa

Presentation Outline:

- Introduction
- Challenges
- Motivations
- Previous versions of WME
    - WME 1.0
    - WME 2.0
- WME 3.0 Building
    - Enhancement of medical concepts
    - Category Assignment for medical concepts
- Evaluation of WME 3.0
    - Validation of overall lexicon
    - Validation of Individual Features
- Conclusions
- References

Introduction:

➢ Information extraction system is essential in healthcare due to the following issues,

- – Structured Corpus Preparation
- – An automated annotation system development
- – Ontology design for medical concepts
- – Medical concepts and its related features extraction
- – Understand the knowledge-based information for the medical corpus

➢ A domain-specific lexicon is important to build an automated information extraction system

Introduction:

➢ A lexicon provides the following information to recognize the contextual knowledge from the medical corpus

- Medical concepts (e.g. Abdomen, Mass)
- Linguistic Features
  - Category
  - Parts-Of-Speech (POS)
  - Gloss (Descriptive Defination)
- Conceptual Features
  - Affinity score (relation between a pair of concepts)
  - Gravity score (relation between concept and its gloss)
- Sense-based Features
  - Polarity score
  - Sentiment
  - Similar Sentiment Words (SSW)

Challenegs:

➢ Unavailability of structured corpora
➢ It is challenging to find an annotated dataset, which combinedly labels the fundamental categories of medical concepts such as Diseases, Symptoms, and Drugs
➢ Isolation of general concepts and medical concepts are difficult
➢ Disambiguation of polarities of the medical concepts
➢ Hard to recognize similar types of diseases or symptoms of a particular disease or symptom

Motivation:

- Development of Structured Corpus
- Medical concepts and related information extraction
- Enrichment of our previously developed medical lexicon (WME 2.0):
  - Enhance more number of medical concepts
  - Recognize the existing features of WME 2.0 for the additional medical concepts
  - Additional category assignment for medical concepts

Motivations:

15 Feb 2000   Pathology Report
Requested By: Todd, Leslie   (GT69-3109)
TISSUE DESCRIPTION:
Tissue from the left brain (0/-10; 1.3 x 1.2 x 0.4 cm in aggregate)
SE03-6829 L9
DIAGNOSIS:
Brain, left frontal, stereotactic biopsy 0/-10: Grade 4 (of 4) small cell astrocytoma.

(Seen with Kim J. Turner and Lee B. Sayer)

Immuno histochemical stains were performed on paraffin sections. Neoplastic cells stain with antibodies to GFAP and to S-100 protein focally. These findings support the above diagnosis.

17 Feb 2000   Green, Cody

PRELIMINARY FROZEN SECTION CONSULTATION
Brain, left, stereotactic biopsy 0/-10: Grade 3

28-Feb-2000 08:07:00   RADIOLOGY REPORT   Exam: MRI Hd wo&w   NORM
Indications: f/u tumor
ORIGINAL REPORT - 28-Feb-2000 12:38:00
MRI of the head without and with gadolinium demonstrates an enhancing, partially cystic, partially solid mass within the posterior left frontal lobe measuring approximately 5.5x4x4cm in dimension which has been shown by biopsy to represent high-grade glioma. Also noted is an approximately 3.7x2.5x2.2cm similar-appearing mass in the anterior right temporal lobe and a more infiltrative-appearing mass involving the right thalamus, measuring approximately 2.2x2x1cm which demonstrates bilobed near-solid enhancement. There is non-enhancing increased T2 signal within the right frontal operculum which appears worrisome for a developing lesion. There are a few patchy areas of satellite enhancement seen lateral to the right temporal lobe lesion, best seen on the coronal images. Overall findings are consistent with multicentric glioma. The right thalamic lesion appears slightly more prominent than on prior CT head studies from 2-15-00 and 2-14-00. The amount of mass effect and edema associated with the left frontal and right temporal lesions appears to have decreased slightly in the interim. There is still residual mass effect in the left frontal lobe resulting in approximately 4mm shift of anterior midline structures from left to right. Partial effacement of the frontal horn of the left lateral ventricle. The right thalamic lesion results in mild mass effect upon the body of the right lateral ventricle. The right temporal lobe lesion results in localized mass effect with effacement of multiple right temporal lobe sulci. The basilar cisterns remain patent. Postoperative changes of left frontal stereotaxic biopsy.

8.544 502.998
d by:
1547 28-Feb-2000 12:38

RX PRESCRIPTION   AGE
                          DATE
NAME
ADRESS
Take Two Pills
a day

<?xml version="3.0" encoding="UTF-8"?>
- <Medical Concepts>
 - <Concept>
    ....
  </Concept>
- <Concept>
   <Title>amnesia</Title>
  - <Properties>
     <POS>noun</POS>
     <Category>disease</Category>
     <Gloss>Loss of memory sometimes including the memory of personal identity due to brain injury, shock, fatigue, repression, or illness or sometimes induced by anesthesia.</Gloss>
      - <SSW and Affinity score>
             blackout (0.674)
             memory loss (0.534)
             stupor (0.429)
             fugue (0.345)
      </SSW and Affinity score>
     <Polarity score> - 0.375</Polarity score>
     <Gravity score>0.170</Gravity score>
     <Sentiment>negative</Sentiment>
   </Properties>
  </Concept>
- <Concept>
    ....
  </Concept>
</Medical Concepts>

Motivation:

```xml
<?xml version="3.0" encoding="UTF-8"?>
- <Medical Concepts>
  - <Concept>
    ........
  </Concept>
- <Concept>
    <Title>amnesia</Title>
  - <Properties>
      <POS>noun</POS>
      <Category>disease</Category>
      <Gloss>Loss of memory sometimes including the memory of
        personal identity due to brain injury, shock, fatigue, repression,
        or illness or sometimes induced by anesthesia.</Gloss>
      - <SSW and Affinity score>
              blackout (0.674)
              memory loss (0.534)
              stupor (0.429)
              fugue (0.345)
        </SSW and Affinity score>
      <Polarity score> - 0.375</Polarity score>
      <Gravity score>0.170</Gravity score>
      <Sentiment>negative</Sentiment>
    </Properties>
  </Concept>
- <Concept>
    ........
  </Concept>
</Medical Concepts>
```

Disease
Hundreds of foods and plant-based fiber products are available to relieve constipation naturally .

Human_anatomy
sounding as if the   nose   were pinched .

Symptom   Disease   Human_anatomy   Disease   Disease
abnormal dryness of the conjunctiva and   cornea   of the eyes; may be due to a systemic deficiency of vitamin A .

Disease   Disease
Giant cell interstitial pneumonia (GIP) is a rare form of pulmonary fibrosis .

Annotated Corpus

Previous versions of WME

Seed list and Useful resources for the previous versions of WME:

➢SemEval 2015, Task-6 Trial and Training Datasets
➢Pre-processed English Medical Dictionary
➢Conventional WordNet
➢SentiWordNet
➢SenticNet
➢Bing Liu subjective list
➢Taboada's adjective list

WME 1.0

"Lexical Resource for Medical Events: A Polarity Based Approach". A Mondal, I Chaturvedi, D Das, R Bajpai, S Bandyopadhyay. 2015. IEEE International Conference on Data Mining Workshop (ICDMW), 1302-1309.

WME 1.0 Lexicon:

➢ Total number of medical concepts: 6415
➢ Parts-Of-Speech viz. Noun, Verb, Adjective etc.
➢ Gloss: Descriptive definition
➢ Polarity score: ranges from -1 to +1
➢ Sentiment: positive or negative

Sample output of WME 1.0 Lexicon:

```xml
<?xml version="1.0" encoding="UTF-8"?>
- <Medical Concepts>
  - <Concept>
      ........
  </Concept>
  - <Concept>
     <Title>abdominal_cavity</Title>
     - <Properties>
        <POS>noun</POS>
        <Gloss>The cavity containing the major viscera; in mammals
         it is separated from the thorax by the diaphragm.</Gloss>
        <Polarity score> - 0.500</Polarity score>
        <Sentiment>negative</Sentiment>
     </Properties>
  </Concept>
  - <Concept>
      ........
  </Concept>
</Medical Concepts>
```

"Lexical Resource for Medical Events: A Polarity Based Approach". A Mondal, I Chaturvedi, D Das, R Bajpai, S Bandyopadhyay. 2015. IEEE International Conference on Data Mining Workshop (ICDMW), 1302-1309.

WME 2.0

1. "WME: Sense, Polarity and Affinity based Concept Resource for Medical Events". A Mondal, D Das, E Cambria, S Bandyopadhyay. 2016. Proceedings of the Eighth Global WordNet Conference, 242-246.

2. "Employing Sentiment-based Affinity and Gravity Scores to Identify Relations of Medical Concepts". A Mondal, E Cambria, D Das, S Bandyopadhyay. 2017. Proceedings of the IEEE Symposium Series on Computational Intelligence (IEEE SSCI 2017) Conference, Honolulu, Hawaii, USA.

WME 2.0 Lexicon:

➢ Total number of medical concepts: 6415
➢ Parts-Of-Speech viz. Noun, Verb, Adjective etc.
➢ Gloss: Descriptive definition
➢ Polarity score: ranges from -1 to +1
➢ Sentiment: positive or negative
➢ SSW: Similar Sentiment Words
➢ Affinity score: ranges from 0 to +1
➢ Gravity score: ranges from -1 to +1

Sentiment based Relational Features:

➢ Sentiment based relations help to identify the hidden links between medical concepts

➢ It also assists in recognizing the proper link between various concepts and their different source of glosses

➢ The proposed sentiment based relational features are Affinity score and Gravity score

Affinity Score:

➢ Affinity refers to the linking between pair of medical concepts by determining sentiment from their common Similar Sentiment Words (SSW)

➢ Affinity score is obtained by a probabilistic count of similar sentiment-based concepts as shown in Equation 1 and 2,

First, we define the overlapping SSW of each concept pair as

$$\text{Affinity}_c = MC_1 \cap MC_2 \qquad (1)$$

where MC_1 and MC_2 represents SSW sets of two different medical concepts and the Affinity_c implies the number of common SSW of MC_1 and MC_2.

Finally, Affinity score for a concept (MC_1) is

$$\text{Affinity Score}_c = \text{Affinity}_c / (MC_1 + MC_2) \qquad (2)$$

Affinity Score:

➢ (a) 4 common SSW *breathing*, *respiration*, *ventilation*, and *external respiration* are determined for the concept pair of *abdominal breathing* and *hypopnea* out of total 8 SSW, resulting in affinity score Affinity Score$_c$ = 0.500

➢ (b) 3 common SSW *breathing*, *ventilation*, and *external respiration* out of total 12 SSW with respect to the pair *abdominal breathing* and *wheeze* and it provides Affinity Score$_c$ = 0.250 that indicates finite but weaker relations of the pair in (a).
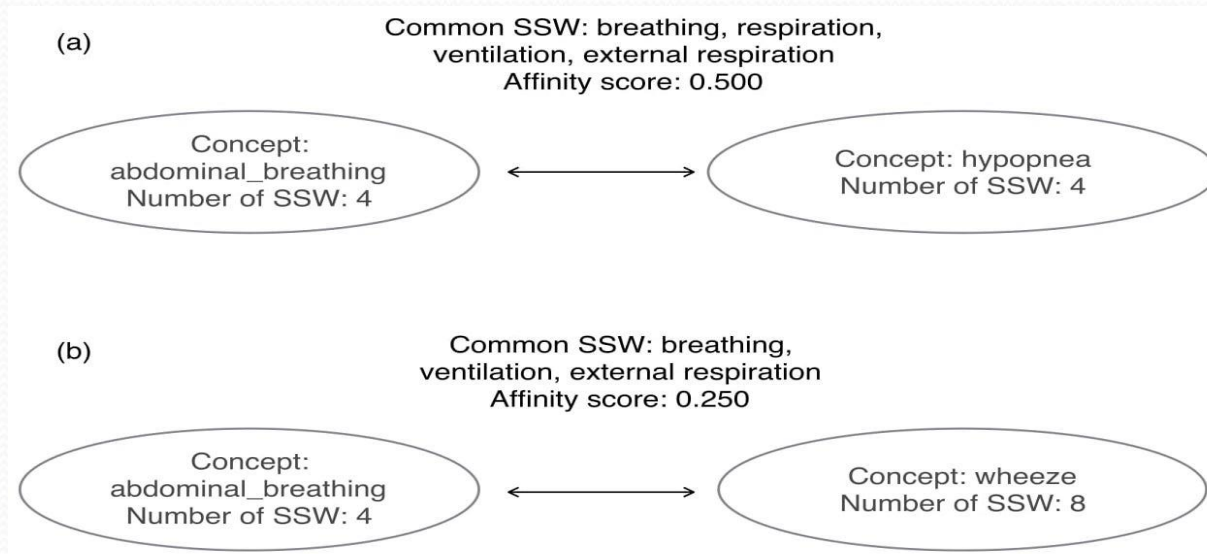
Gravity Score:

➢Gravity presents the relevance of sentiment appeared between a medical concept and its glosses.

➢Gravity score confirms whether a gloss (or contextual information) pertaining a medical concept appropriates or not.

In order to achieve gravity score, we first consider the polarity score of each word in glosses and is denoted as $Polarity_{gc}$.

Secondly, we calculate the gravity score $Gravity_{gc}$ for each medical concept which aims to determine the sentiment relations with the attached gloss by considering the polarity of medical concepts $Polarity_c$.
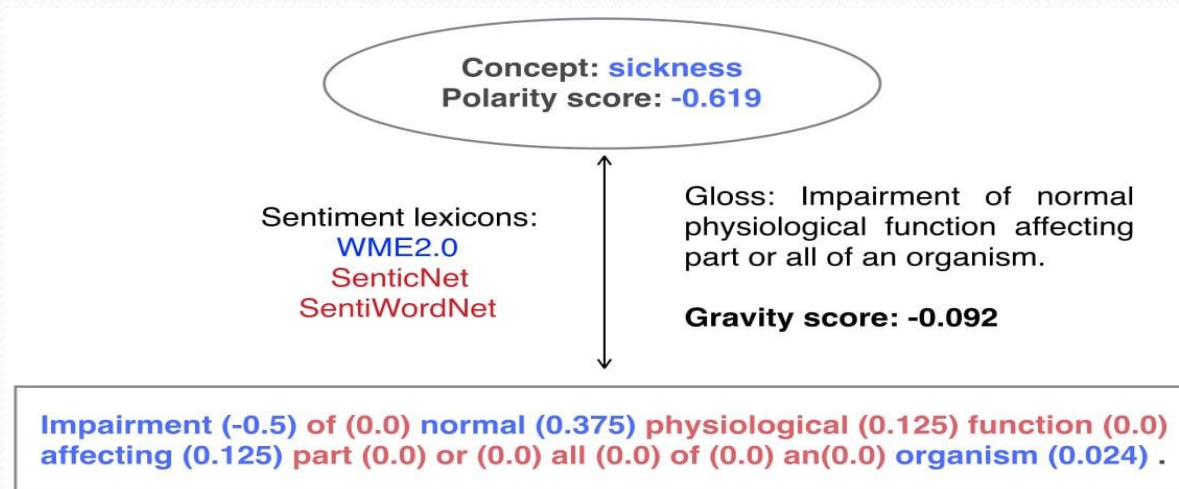
$$Gravity_{gc} = Polarity_c \times Polarity_{gc} \qquad (3)$$

The final gravity score is then simply as,

$$Gravity\ Score_c = \sum Gravity_{gc}, i=1\ to\ N \quad (4)$$

Gravity Score:

➤ The medical concept *sickness* with the polarity score $Polarity_c = -0.619$ is given with its gloss along with the medical concepts of WME 2.0 and represented in blue color whereas all others are shown as red.

➤ Every word is provided with its gross polarity $Polarity_{gc}$ as mentioned in the parenthesis.

➤ The score is assigned by the corresponding lexicons, e.g., the scores of medical concepts are collected from WME 2.0 and the others from either SenticNet or SentiWordNet.

Concept: sickness
Polarity score: -0.619

Sentiment lexicons:
WME2.0
SenticNet
SentiWordNet

Gloss: Impairment of normal physiological function affecting part or all of an organism.

Gravity score: -0.092

Impairment (-0.5) of (0.0) normal (0.375) physiological (0.125) function (0.0) affecting (0.125) part (0.0) or (0.0) all (0.0) of (0.0) an(0.0) organism (0.024) .

Sample output of WME 2.0 Lexicon:

```xml
<?xml version="2.0" encoding="UTF-8"?>
- <Medical Concepts>
 - <Concept>
    ........
   </Concept>
- <Concept>
   <Title>amnesia</Title>
   - <Properties>
     <POS>noun</POS>
     <Gloss>Loss of memory sometimes including the memory of
      personal identity due to brain injury, shock, fatigue, repression,
      or illness or sometimes induced by anesthesia.</Gloss>
      - <SSW and Affinity score>
              blackout (0.674)
              memory loss (0.534)
              stupor (0.429)
              fugue (0.345)
        </SSW and Affinity score>
     <Polarity score> - 0.375</Polarity score>
     <Gravity score>0.170</Gravity score>
     <Sentiment>negative</Sentiment>
    </Properties>
  </Concept>
- <Concept>
    ........
  </Concept>
</Medical Concepts>
```

"WME: Sense, Polarity and Affinity based Concept Resource for Medical Events". A Mondal, D Das, E Cambria, S Bandyopadhyay. 2016. Proceedings of the Eighth Global WordNet Conference, 242-246.

WME 3.0 Building

WME 3.0 Lexicon:

➢ Medical ontology based resource preparation
➢ Total number of medical concepts: 10186
➢ Parts-Of-Speech viz. Noun, Verb, Adjective etc.
➢ Gloss: Descriptive definition
➢ Polarity score: ranges from -1 to +1
➢ Sentiment: positive or negative
➢ Category: Diseases, Symptoms, Drugs, Human Anatomy and MMT (Miscellaneous Medical Terms)
➢ SSW: Similar Sentiment Words
➢ Affinity score: ranges from 0 to +1
➢ Gravity score: ranges from -1 to +1

Additional Medical Concepts Identification:

➢ Two resources namely WordNet and MedicineNet assist in enhancing 3771 number of medical concepts with WME 2.0

➢ Recognize features viz. POS, gloss, SSW, polarity score, sentiment, affinity score and gravity score for these additional concepts using following resources and machine learning approaches
  - ➢ SentiWordNet
  - ➢ SenticNet
  - ➢ Bing Liu subjective list
  - ➢ Taboada's adjective list
  - ➢ A preprocessed medical dictionary

Category Assignment to Medical Concepts:

➢ Assigned categories are diseases, symptoms, drugs, human anatomy and Miscellaneous Medical Terms (MMT)

➢ WME 3.0 assigned features of medical concepts help to assign the category of medical concepts in the presence of two well-known classifiers

➢ The classifiers are Naive Bayes and Logistic Regression

Sample output of WME 3.0 Lexicon:

```xml
<?xml version="3.0" encoding="UTF-8"?>
- <Medical Concepts>
  - <Concept>

      ........
  </Concept>
- <Concept>
    <Title>amnesia</Title>
  - <Properties>
      <POS>noun</POS>
      <Category>disease</Category>
      <Gloss>Loss of memory sometimes including the memory of
       personal identity due to brain injury, shock, fatigue, repression,
       or illness or sometimes induced by anesthesia.</Gloss>
         - <SSW and Affinity score>
                    blackout (0.674)
                    memory loss (0.534)
                    stupor (0.429)
                    fugue (0.345)
           </SSW and Affinity score>
        <Polarity score> - 0.375</Polarity score>
        <Gravity score>0.170</Gravity score>
        <Sentiment>negative</Sentiment>
    </Properties>
  </Concept>
- <Concept>

      ........
  </Concept>
</Medical Concepts>
```

A comparative study between all three versions of WME:

		WME 1.0	WME 2.0	WME 3.0
No. of Concepts		6415	6415	10186
n-grams	**Uni-gram**	2956	2956	3722
	Bi-gram	2837	2837	3866
	Tri-gram	622	622	1762
POS	**Nouns**	4248	4248	7677
	Verbs	2056	2056	2352
	Adjectives	111	111	157
Sentiment and Polarity score	**Positive (> = 1)**	2800	2800	3227
	Negative (< 1)	3615	3615	6959

		WME 1.0	WME 2.0	WME 3.0
Affinity score	**0 to 0.5**	-	4325	7177
	0.5 to 1	-	2090	3009
Gravity score	**Less than zero**	-	2320	3783
	Equal to zero	-	732	1961
	Grater than zero	-	3363	4442
Category	**Diseases**	-	-	3243
	Drugs	-	-	3390
	Symptoms	-	-	1409
	Human anatomy	-	-	227
	MMT	-	-	1917

Evaluation

Evaluation Process:

➢We have conducted an agreement analysis with the help of two manual annotators to evaluate WME 3.0 lexicon
➢The manual annotators are medical practitioners
➢The agreement score (k score) has been calculated using the following equation,

$$\kappa = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)}$$

where, k is the cohen's kappa agreement score
Pr(a) is the observed proportion of full agreement between two annotators.
Pr(e) is the proportion expected by a chance which indicates a kind of random agreement between the annotators.

Validation of overall WME 3.0 lexicon:

No. of Concepts: 10186		Annotator-1	
		Yes	No
Annotator-2	Yes	8629	189
	No	285	1083

Cohen's Kappa score (k) = 0.79

Validation of individual features of WME 3.0:

No. of Concepts: 10186			Annotator-1		K score
			Yes	No	
Annotator-2	Category	Yes	8778	93	0.89
		No	161	1154	
	POS	Yes	9229	52	0.91
		No	92	813	
	Gloss	Yes	8805	97	0.88
		No	172	1112	

No. of Concepts: 10186			Annotator-1		K score
			Yes	No	
Annotator-2	SSW	Yes	8767	137	0.82
		No	256	1026	
	Sentiment	Yes	8727	67	0.92
		No	124	1268	

Validation of individual categories of WME 3.0:

No. of Concepts: 10186			Annotator-1		K score
			Yes	No	
Annotator-2	Disease (3243)	Yes	2794	31	0.89
		No	51	367	
	Symptom (1409)	Yes	1214	14	0.87
		No	26	155	
	Drug (3390)	Yes	2922	34	0.88
		No	53	381	

Continue..

No. of Concepts: 10186			Annotator-1		K score
			Yes	No	
Annotator-2	Human anatomy (227)	Yes	196	2	0.90
		No	3	26	
	MMT (1917)	Yes	1652	12	0.91
		No	28	225	

Conclusions

Conclusions:

➢ An enriched version of a medical lexicon viz. WME 3.0

➢ Category assignment for Medical concepts

➢ Various syntactic and semantic features extraction for medical concepts

➢ WME 3.0 may help to reduce the gap between medical experts and non-experts

➢ Assistance to prepare an annotated corpus

References

References:

[1] Anupam Mondal, Iti Chaturvedi, Dipankar Das, Rajiv Bajpai, and Sivaji Bandyopadhyay. 2015. Lexical Resource for Medical Events: A Polarity Based Approach. In 2015 IEEE International Conference on Data Mining Workshop (ICDMW), pages 1302–1309. IEEE.

[2] Anupam Mondal, Dipankar Das, Erik Cambria, and Sivaji Bandyopadhyay. 2016. WME: Sense, Polarity and Affinity based Concept Resource for Medical Events. Proceedings of the Eighth Global WordNet Conference, pages 242–246.

[3] Anupam Mondal, Erik Cambria, Dipankar Das, and Sivaji Bandyopadhyay. 2017a. Auto-categorization of medical concepts and contexts. In IEEE Symposium Series on Computational Intelligence (SSCI 2017), Honolulu, Hawaii, USA.

[4] Anupam Mondal, Erik Cambria, Dipankar Das, and Sivaji Bandyopadhyay. 2017b. MediConceptNet: An Affinity Score Based Medical Concept Network. In Proceedings of the Thirtieth International Florida Artificial Intelligence Research Society Conference, FLAIRS 2017, Marco Island, Florida, USA, May 22-24, 2017., pages 335–340.

[5] Barry Smith and Christiane Fellbaum. 2004. Medical WordNet: a new methodology for the construction and validation of information resources for consumer health. In Proceedings of the 20th international conference on Computational Linguistics, page 371. Association for Computational Linguistics.

References:

[6] Erik Cambria and Amir Hussain. 2015. Sentic Computing: A Common-Sense-Based Framework for Concept-Level Sentiment Analysis. Springer, Cham, Switzerland.

[7] Erik Cambria, Amir Hussain, Tariq Durrani, Catherine Havasi, Chris Eckl, and James Munro. 2010a. Sentic computing for patient centered applications. In IEEE 10th International Conference on Signal Processing Proceedings, pages 1279–1282. IEEE.

[8] Christiane Fellbaum. 1998. WordNet. Wiley Online Library.

[9] Bing Liu. 2012. Sentiment analysis and opinion mining. Synthesis lectures on human language technologies, 5(1):1–167.

[10] Andrea Esuli and Fabrizio Sebastiani. 2006. SentiWordNet: A publicly available lexical resource for opinion mining. In Proceedings of LREC, volume 6, pages 417–422. Citeseer.

Thank You....