A Hybrid Approach based Sentiment Extraction from Medical Contexts

[1]Anupam Mondal [2]Ranjan Satapathy [1]Dipankar Das [1]Sivaji Bandyopadhyay
[1]Computer Science and Engineering, Jadavpur University, India
[1]anupam@sentic.net, [1]ddas@cse.jdvu.ac.in, [1]sbandyopadhyay@cse.jdvu.ac.in
[2]School of Computer and Information Sciences, University of Hyderabad, India
[2]kumarsatpathy@gmail.com

Abstract

In the domain of Bio medical Natural Language Processing (Bio-NLP), the information extraction and context sentiment identification are treated as emerging tasks. Several linguistic features like negation, uni-gram, bi-gram, Part-of-Speech (POS) have been used to extract the medical concepts and their sense-based context level information. Thus, in the present attempt, a hybrid approach which is the combination of both linguistic and machine learning approaches has been introduced to extract the contextual sense-based information from a medical corpus. The extraction of sentiment oriented keywords is the crucial part towards identifying the senses of medical contexts. In our previous work, we have developed a medical sense-based lexicon known as WordNet of Medical Event (WME). Several sentiment lexicons like SentiWordNet, SenticNet etc. were used to represent WME. In contrast, one of our primary motivations here is to build a sentiment extraction model based on medical contexts to leverage the knowledge of WME using a hybrid approach. The developed model is based on two phases, namely preprocessing phase and learning phase. The preprocessing phase is responsible for extracting and preparing structural data from the raw contexts whereas the learning phase helps to identify the sentiment patterns and evaluate the sentiment extraction process. The two phased hybrid model provides us 81% accuracy for extracting the sentiment based medical contexts as positive and negative by employing NaïveBayes and Sequential minimal optimization (SMO) supervised classifiers.

1 Introduction

One of the major objectives of Sentiment Analysis is to identify and extract the subjective information from a given text using rule based or machine learning approaches [Cambria, 2016]. The domain specific knowledge with above mentioned approaches help us to extract the contextual sentiment information from the medical corpus. Due to lack of involvement of domain experts and unavailability of domain specific structured corpus, the task is challenging in Bio-NLP domain. To overcome the scarcity of such domain specific knowledge for sentiment analysis, several lexicons have been developed like Medical Event Net (MEN), Medical Fact Net (MFN), Medical Belief Net (MBN) and WordNet of Medical Event (WME) [Cambria *et al*., 2010]. These lexicons help to extract the sense of a medical concept, fact and belief oriented information. The present paper reports the development of a medical context based sentiment extraction model. Hence, one of our primary aims is to identify the sense-based concepts from the medical contexts and extract their related sentiment features. In order to identify the sense-based medical concepts, we have introduced the current version of WordNet of Medical Event (WME2.0) knowledge base. WME2.0 contains the medical concept information with their related linguistic and sense-oriented features like POS, gloss of the concept, semantics, polarity score, affinity score, gravity score and sense(s). Among all these features, we have only considered the sense-based features like semantics, polarity score, affinity score and sense to develop our present sentiment extraction model [Swaminathan *et al*., 2010]. On the top of extracted medical concepts based on WME2.0 lexicon, we have applied linguistic and machine learning approaches to get the final sentiment of the contexts. The linguistic approach helps to manage the negation of the contexts as well as derive new rules to extract the sense(s) of such contexts. The POS, unigram, bi-gram, affinity score, polarity score and sense features of the medical concepts of WME2.0 help to extract the sentiment of the medical contexts. The supervised machine learning approach has been introduced to verify the contextual sentiment extracted using linguistic approach. In the process, we have applied NaïveBayes and Sequential minimal optimization (SMO) supervised machine learning classifiers on the derived linguistic features.

In the paper, we have incorporated both linguistic and machine learning approaches together as a hybrid model to leverage the sentiment oriented knowledge of both the domain [Villena-Romn *et al*., 2011]. The proposed hybrid model follows two phase architecture namely preprocessing phase and learning phase. In pre-processing phase, we have focused on the preparation of structured medical concepts from the raw medical contexts and the

Proceedings of the 4th Workshop on Sentiment Analysis where AI meets Psychology (SAAIP 2016), IJCAI 2016, pages 35-40, New York City, USA, July 10, 2016.

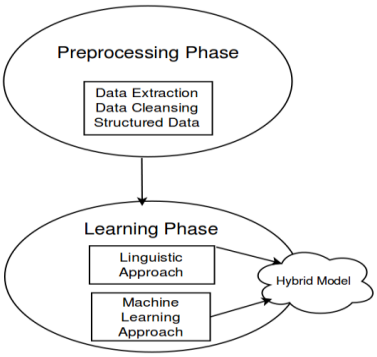
learning phase helps to extract the sentiment of such contexts and evaluate them. The two phase model generates the output in the form of positive or negative sentiment of the context. The hybrid approach based learning phase provides 81% accuracy to extract the medical context based sentiment information.

The remainder of the paper is structured as follows, Section 2 presents related work followed by model design describing the pre-processing and learning phases in Section 3. Section 4 talks about the model discussion and evaluation process we have followed in the paper. Finally, in Section 5, we present our conclusion and future scopes of the model.

2 Related Work

Sentiment analysis of medical contexts is contributory and growing research field under Bio-NLP domain [Cambria *et al*., 2013]. A large number of unstructured corpora and lack of domain experts' involvement have introduced more challenge in this task. In the process, the researchers focused on developing medical sentiment-based lexicon to identify the sentiments of medical concepts. Therefore, the medical concepts and their sense based features indeed help to identify the sentiment of the medical contexts. The linguistic, machine learning and hybrid approaches have been introduced to build the concept and context based sentiment extraction systems. The linguistic approach helps to find the negation words, phrases and construct the knowledge-based rules (with unigram, bigram and n-gram features) for the context level sentiment extraction [Elkin *et al*., 2005; Niu *et al*., 2005; Szarvas *et al*., 2008]. Smith and Fellbaum, 2004 developed a Medical Word-Net (MEN) along with two subnetworks, namely Medical FactNet (MFN) and Medical BeliefNet (MBN), for the evaluation of consumer health reports [Smith and Fellbaum, 2004]. MEN was developed with the help of formal architecture of the Princeton WordNet [Fellbaum, 1998]. MFN serves to assist the non-expert group in providing a better understanding of basic medical information. MBN identifies beliefs about the medical phenomenon. Their primary motivation was to develop a network of medical information retrieval systems with visualization effect. The domain-specific knowledge and the abovementioned features are essential to improve the efficiency of the sentiment extraction system [Shukla *et al*., 2015]. So, these approaches were not able to provide adequate accuracy due to the lack of knowledge involvement from the domain experts. Hence, to overcome the mentioned problem, the researchers introduced supervised machine learning approaches [Smith and Lee, 2012]. Standard NaïveBayes, Multinomial NaïveBayes and Support Vector Machine (SVM) supervised classifiers were applied with unigram, bigram, Parts Of Speech (POS) and negation features under the machine learning framework. The researchers have also used hybrid approaches to improve the accuracy of the medical context based sentiment extraction systems. One of the hybrid approaches was developed with the combination of linguistic and machine learning approaches [Boytcheva *et al*., 2005; Villena-Romn *et al*., 2011]. Sohn *et al*., 2012, developed an emotion identification system from suicide notes using the hybrid approach [Sohn *et al*., 2012]. The suicide notes were provided by the challenge organizers of Informatics for Integrating Biology and the Bedside (I2B2). Machine learning, linguistic rule-based and their combined approaches have been applied to the training dataset of the suicide notes and the system provided 0.5640 micro-average F-score for the training dataset. Birks et al., 2009, applied the combination of RIPPER (Repeated Incremental Pruning to Produce Error Reduction), multinomial NaïveBayes classifier and manual pattern matching rules to identify the emotions of the sentences [Birks *et al*., 2009]. Mondal et al., 2016, developed WordNet of Medical Events (WME) lexicon to identify the medical concepts and their knowledge-based and semantic features using hybrid approach [Mondal *et al*., 2015]. The latest version of WME (WME2.0) contains POS, semantics, gloss, affinity score, gravity score, polarity score and sense features of the concepts [Mondal *et al*., 2016]. WME2.0 sentiment lexicon has identified the senses of the concepts using SentiWordNet[1], SenticNet[2], BingLiu[3] and Taboda's adjective list [Mondal *et al*., 2016; Mondal *et al*., 2015; Taboada *et al*., 2011]. In this paper, we have used the WME2.0 lexicon to identify the concepts and their features to extract sentiments of the medical contexts.

Figure 1: Two phase proposed Model



3 Model Design

The knowledge-based sentiment lexicon is crucial to design a context based sentiment extraction system. The medical concepts and their linguistic features are extracted from the domain-specific sentiment lexicon. To overcome the problem of experts' availability, we have formulated WME2.0 lexicon with a hybrid approach. It adds an extra dimension

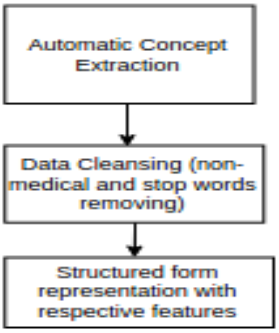
[1] http://sentiwordnet.isti.cnr.it/
[2] http://sentic.net/
[3] https://www.cs.uic.edu/liub/FBS/sentiment-analysis.html

for improving the accuracy of the extracted medical context sentiment. The proposed hybrid approach is the combination of linguistic and machine learning approach. The approach consists of two phases namely pre-processing and learning phase. Figure 1 shows the architecture of the proposed approach (model).

3.1 Pre-processing phase

The phase extracts the sentiments of medical contexts in the form of context related medical concepts, their sentiments and knowledge-based information. The structured form of the concepts is essential in identifying the important medical concepts from the context.

Figure 2: Flowchart of Preprocessing Phase



In this concern, to represent the structured medical concepts, the required steps are data extraction, cleansing and formatting. The research community provided various linguistic resources such as open source data preprocessing tools (viz. NLTK, stemming etc.) [Na *et al*., 2012]. The following steps illustrate the basic operations of the pre-processing phase:

Data Extraction: The medical concepts extraction from a given context is the primary task of this step. WME2.0 helps to extract the medical concepts and their linguistic and sense-based features from the context. Moreover, the non-medical concepts and their sense identification are also essential to identify the sentiment of the context. The non-medical concepts the senses have been extracted using SentiWordNet and SenticNet lexicons [Cambria *et al*., 2014; Cambria *et al*., 2013; Esuli and Sebastiani, 2006].

Data Cleansing: Data cleansing step is responsible to remove the context related stop-words and stemmed the concept words. The classification of medical and nonmedical concepts and identification of negation words (like no, not, never etc.) are also taken care of by data cleansing step [Huang and Lowe, 2007].

Data Formatting: Data formatting has been applied to represent the structured form of the extracted medical concepts [Hussain *et al*., 2011]. The extracted structured (vector) concepts have been forwarded to the learning phase along with their features. The concept structure is represented as follows:

<Concept (gastric), POS (noun), Semantics (abdominal breathing, visceral, intestinal, belly, duodenal, stomachic), Polarity Score (-0.5), Sense (Negative)>

3.2 Learning phase

Followed by the pre-processing phase, the hybrid approach has been introduced in the learning phase to build the contextual sentiment extraction system. Linguistic and machine learning has been combined to form the hybrid approach. The linguistic approach with WME2.0 knowledge base lexicon helps to identify the hidden rules. These rules are able to extract the concept sentiment and their polarity. The extracted linguistic concept features (rules) were fed to the supervised machine learning classifiers to evaluate the accuracy of the model. The linguistic approach provides a support to handle the negation effect of the context and help to identify the appropriate sentiment of the context [Huang and Lowe, 2007]. The learning phase is illustrated as follows:

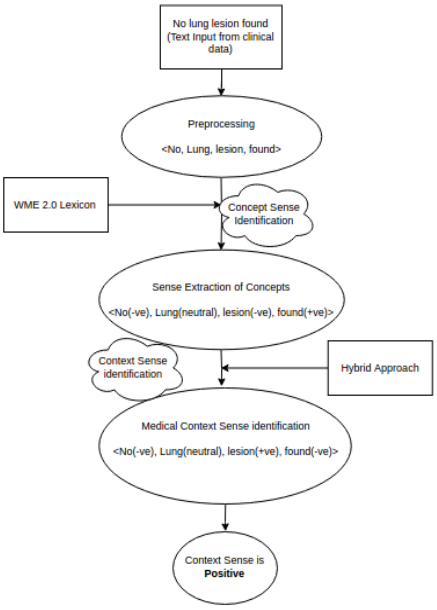
Step 1: Identify the polarity score and sense of each concept (medical and non-medical) of the context.
Step 2: Linguistic approach-based negation words (concept) handling.
Step 3: Calculate the overall polarity of the context.
$\text{Context}_{\text{polarity}} = \sum \text{Polarity}_c$
Where, c = number of concepts in the context and Polarity_c indicates the polarity score of each concept.
Step 4: The context sentiment has been evaluated using $\text{Context}_{\text{polarity}}$ score.

4 Discussion and Evaluation

The context related medical concepts and their semantic features (extraction polarity, semantics and sense) are required to identify the sentiment of the medical context [Sarker *et al*., 2011]. In the process, the statistical and linguistic features based medical sentiment lexicons were facing difficulties due to the unstructured nature of the corpus. So, the researchers tried to build an intelligent automated sentiment extraction system in the Bio-NLP domain [Shukla *et al*., 2015; Sohn *et al*., 2012]. The system helps to extract the structured knowledge-based information with a proper sentiment of the context. WordNet of Medical Event (WME2.0) was introduced to identify the medical concept and their sense-based features. The WME2.0 lexicon able to extract the medical concepts and their POS, semantics, gloss, affinity score, gravity score, polarity score and sense. On the top of WME2.0 lexicon, the hybrid approach has been applied to extract the context level sentiment for the

proposed model. The model is based on two phases namely pre-processing and learning phase. The pre-processing phase has considered the concept extraction (medical and non-medical concept), concept cleansing (concept stemming and stop-words removing) and concept formatting <Concept, POS, semantic, polarity score, sense> steps. The learning phase identified the sentiment using the linguistic and machine learning approaches on the pre-processing step driven data. The concept linguistic features and knowledge based WME sentiment resource help to extract the overall context sentiment and polarity score. The linguistic approach provides a support to handle the negation and identifies the correct sense of the context. The medical context "No lung lesion found" has been evaluated as "positive" sentiment after handling the negation. The system first extracts the concepts and their sense as "no (-ve)", "lung (neutral)", "lesion (-ve)" and "found (+ve)" using WME2.0 resource. The linguistic-based negation handling approach has been applied on the extracted sense and identify the overall context sense as "positive". In the learning phase, the hybrid approach has been introduced to extract and measure the accuracy of the context sentiment. The linguistic approach involves knowledge-based medical concept mapping with WME2.0 lexicon. Further, the NaïveBayes and Sequential minimal optimization (SMO) support vector based supervised machine learning approaches have been employed for evaluating the accuracy of the model. Figure 3 and Figure 4 describe the positive and negative contexts with respect to the sentiment extraction process, respectively.

Figure 3: Positive Sentiment extraction



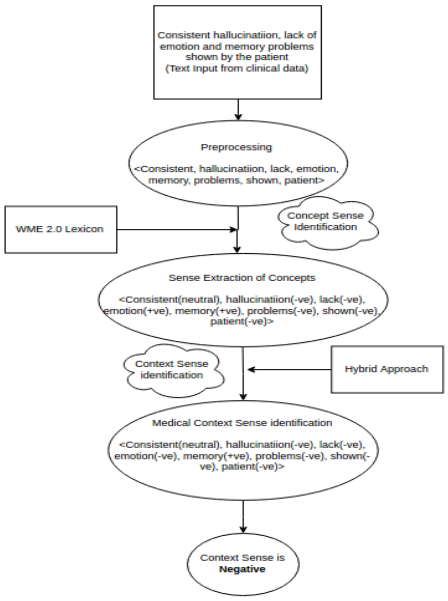
4.1 Evaluation Process

To develop and measure the accuracy of the context level sentiment extraction system, the data has been collected from the open source resource[4]. We have extracted 7042 number of medical contexts and applied through the proposed sentiment extraction system. The context sentiment extraction system has provided 3265 number of the positive and 3777 number of the negative sentiments of the contexts. To evaluate the extracted context sentiment, the linguistic features (number of negation word, context polarity score and sense) were fed to the NaïveBayes and support vector based SMO supervised machine learning classifiers under the WEKA[5] tool. The extracted 7042 number of context data has been represented as 4900 number of training and the remaining 2142 number of test dataset. The system's accuracy was measured as F-Measure with four types of models like, Use training set, Supplied test set, Cross-validation Folds 10 and Percentage split %66. Table 1 shows the F-Measures of these modes for the NaïveBayes and support vector based SMO supervised classifiers. The linguistic and machine learning based hybrid approach provides the accuracy score nearly 81% for the medical context sentiment extraction model.

Table 1: F-Measure of Supervised classifiers

Model	NaïveBayes	SMO
Use training set	0.868	0.890
Supplied test set	0.815	0.815
Cross-validation Folds 10	0.864	0.867
Percentage split %66	0.873	0.879

Figure 4: Negative Sentiment extraction



[5] http://weka.wikispaces.com/

5 Conclusion and Future scope

Sentiment or opinion analysis is important to extract the contextual information from the medical context under NLP domain. The context sentiment helps to identify the knowledge based information and proper utilization of the context. The paper has reported a hybrid approach based context sentiment extraction model with two phases. The phases are preprocessing (important medical keywords extraction) and learning (respective sentiment identification). In the process, the linguistic and machine learning combined hybrid approach has been applied on the top of WordNet of Medical Event (WME2.0) lexicon to extract the medical concepts in order to identify the sentiment of the medical context. The medical concept polarity score and their related sense helps to identify the medical context sentiment [Cambria, 2013] and [Cambria *et al*., 2015]. WME2.0 lexicon driven medical concepts affinity score and their semantic features are crucial in building the proposed model. The medical concept semantics, polarity score and affinity score helps to identify the medical concept sentiment with polarity score. The hybrid approach provides nearly 81% accuracy for the proposed context sentiment extraction system. Hence, the future research will focus to develop some practical applications relating to the current work as medical annotation and context summarization system. These systems will provide the support to the expert and non-expert groups in their respective applications.

References

[Mondal *et al*., 2016] Anupam Mondal, Dipankar Das, Erik Cambria and Sivaji Bandyopadhyay. WME: Sense, polarity and affinity based concept resource for medical events. In *Proceedings of the Eighth Global WordNet Conference*, pages 242–246, 2016.

[Birks *et al*., 2009] Yvonne Birks, Jean McKendree, and Ian Watt. Emotional intelligence and perceived stress in healthcare students: a multi-institutional, multi-professional survey. *BMC Medical Education*, 9(1):1–8, 2009.

[Boytcheva *et al*., 2005] Svetla Boytcheva, Albena Strupchanska, Elena Paskaleva, Dimitar Tcharaktchiev, and Dame Gruev Str. Some aspects of negation processing in electronic health records. In *Proceedings of International Workshop Language and Speech Infrastructure for Information Access in the Balkan Countries*. Pages 1—8, 2005.

[Cambria *et al*., 2010] E. Cambria, A. Hussain, T. Durrani, C. Havasi, C. Eckl, and J. Munro. Sentic computing for patient centered applications. In *IEEE 10th International Conference on Signal Processing Proceedings*, pages 1279–1282, Oct 2010.

[Cambria, 2013] Erik Cambria. An introduction to concept-level sentiment analysis. In *Advances in Soft Computing and Its Applications - 12th Mexican International Conference on Artificial Intelligence*, MICAI 2013, Mexico City, Mexico, November 24-30, 2013, Proceedings, Part II, pages 478–483, 2013.

[Cambria, 2016] Erik Cambria. Affective computing and sentiment analysis. *IEEE Intelligent Systems*, 31(2):102–107, 2016.

[Cambria *et al*., 2015] Erik Cambria, Jie Fu, Federica Bisio, and Soujanya Poria. Affectivespace 2: Enabling affective intuition for concept-level sentiment analysis. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, January 25-30, 2015, Austin, Texas, USA, pages 508–514, 2015.

[Cambria *et al*., 2014] Erik Cambria, Daniel Olsher, and Dheeraj Rajagopal. Senticnet 3: A common and common-sense knowledge base for cognition-driven sentiment analysis. In *AAAI Conference on Artificial Intelligence*, 2014.

[Cambria *et al*., 2013] Erik Cambria, Bjrn Schuller, Yunqing Xia, and Catherine Havasi. New avenues in opinion mining and sentiment analysis. *IEEE Intelligent Systems*, 28(2):15–21, 2013.

[Hussain *et al*., 2011] Hussain A Cambria E and Eckl C. Bridging the gap between structured and unstructured health- care data through semantics and sentics. In *Proceedings of ACM WebSci*, Koblenz, 2011.

[Elkin *et al*., 2005] Peter L. Elkin, Steven H. Brown, Brent A. Bauer, Casey S. Husser, William Carruth, Larry R. Bergstrom and Dietlind L. Wahner-Roedler. A controlled trial of automated classification of negation from clinical notes. *BMC Medical Informatics and Decision Making*, 5(1):1–7, 2005.

[Esuli and Sebastiani, 2006] Andrea Esuli and Fabrizio Sebastiani. Sentiwordnet: A publicly available lexical resource for opinion mining. In *Proceedings of the 5th Conference on Language Resources and Evaluation (LREC06)*, pages 417–422, 2006.

[Fellbaum, 1998] Christiane Fellbaum. WordNet: an electronic lexical database. *MIT Press*, 1998.

[Huang and Lowe, 2007] Yang Huang and Henry J. Lowe. A novel hybrid approach to automated negation detection in clinical radiology reports. *Journal of the American Medical Informatics Association: JAMIA*, 14(3):304–311, May 2007.

[Mondal *et al*., 2015] Anupam Mondal, Iti Chaturvedi, Dipankar Das, Rajiv Bajpai, and Sivaji Bandyopadhyay. Lexical resource for medical events: A polarity based approach. In *IEEE ICDM Workshops*, pages 1302–1309. IEEE, 2015.

[Na *et al*., 2012] Jin-Cheon Na, Wai Yan Min Kyaing, Christopher SG Khoo, Schubert Foo, Yun-Ke Chang, and Yin-Leng Theng. Sentiment classification of drug reviews using a rule-based linguistic approach. In *The outreach of digital libraries: a globalized resource network*, pages 189–198. Springer, 2012.

[Niu *et al*., 2005] Yun Niu, Xiaodan Zhu, Jianhua Li, and Graeme Hirst. Analysis of polarity information in medical text. In *Proceedings of the American Medical Informatics Association Annual Symposium*, 2005.

[Sarker *et al*., 2011] Abeed Sarker, Diego Moll´a-Aliod, C´ecile Paris, et al. Outcome polarity identification of medical papers. *Melbourne: Australian Language Technology Association*. 2011.

[Shukla *et al*., 2015] Ravi Shankar Shukla, Kamendra Singh Yadav, Syed Tarif Abbas Rizvi, and Faisal Haseen. An Efficient Mining of Biomedical Data from Hypertext Documents via NLP. In *Proceedings of the 3rd International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA) 2014*: Volume 1, pages 651–658. Springer International Publishing, Cham, 2015.

[Smith and Fellbaum, 2004] Barry Smith and Christiane Fellbaum. Medical wordnet: A new methodology for the construction and validation of information resources for consumer health. In *Proceedings of COLING*, 2004.

[Smith and Lee, 2012] Phillip Smith and Mark Lee. Cross-discourse development of supervised sentiment analysis in the clinical domain. In *Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis, WASSA '12*, Association for Computational Linguistics, pages 79–83, Stroudsburg, PA, USA, 2012.

[Sohn *et al*., 2012] Sunghwan Sohn, Manabu Torii, Dingcheng Li, Stephen Wu, Hongfang Liu, and Avishwar Wagholikar. A Hybrid Approach to Sentiment Sentence Classification in Suicide Notes. In *Biomedical Informatics Insights*, pages 43+, January 2012.

[Swaminathan *et al*., 2010] Rajesh Swaminathan, Abhishek Sharma, and Hui Yang. Opinion mining for biomedical text data: Feature space design and feature selection. In *The Nineth International Workshop on Data Mining in Bioinformatics, BIOKDD*, 2010.

[Szarvas *et al*., 2008] Gy¨orgy Szarvas, Veronika Vincze, Rich´ard Farkas, and J´anos Csirik. The bioscope corpus: annotation for negation, uncertainty and their scope in biomedical texts. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, Association for Computational Linguistics, pages 38–45, Columbus, Ohio, June 2008.

[Taboada *et al*., 2011] Maite Taboada, Milan Tofiloski, Julian Brooke, Kimberly Voll, and Manfred Stede. Lexicon-based methods for sentiment analysis. Journal of *Computational linguistics*, volume 37, number 2, pages 267-307, publisher MIT Press, 2011.

[Villena-Romn *et al*., 2011] Julio Villena-Romn, Sonia Collada-Prez, Sara Lana-Serrano, and Jos Carlos Gonzlez Cristbal. Hybrid approach combining machine learning and a rule-based expert system for text categorization. In *R. Charles Murray and Philip M. McCarthy, editors, FLAIRS Conference*. AAAI Press, 2011.