Procesamiento del Lenguaje Natural, Revista nº 58, marzo de 2017, pp. 101-108 recibido 30-11-2016 revisado 13-01-2017 aceptado 07-02-2017

Analysis of patient satisfaction in Dutch and Spanish online reviews

Análisis de la satisfacción del paciente a partir de comentarios online escritos en holandés y en español

Salud María Jiménez-Zafra,
M. Teresa Martín-Valdivia,
Sinai Group
Universidad de Jaén
Campus Las Lagunillas s/n. E-23071
{sjzafra, maite}@ujaen.es

Isa Maks,
Rubén Izquierdo
Computational Linguistics
VU University Amsterdam
De Boelelaan 1105, 1081 HV
{isa.maks,ruben.izquierdobevia}@vu.nl

Abstract: Sentiment Analysis is a well-known task of Natural Language Processing that has been studied in different domains such as movies, phones or hotels. However, other areas like medical domain remain yet unexplored. In this paper we study different polarity classification techniques applied on health domain. We present a corpus of patient reviews composed by a Dutch part (COPOD: Corpus of Patient Opinions in Dutch) and a Spanish part (COPOS: Corpus of Patient Opinions in Spanish). Experiments have been carried out using a supervised method (SVM), a cross-domain method (OpeNER) and a dictionary lookup method for both languages. Obtained results overcome the baseline in almost all the cases and are higher than other polarity classifiers in patient domain. Regarding the bilingualism, the developed systems for Dutch and Spanish have a similar performance for F1-measure and Accuracy.
Keywords: Polarity classification, medical domain, patient opinion corpus, opener

Resumen: El Análisis de Sentimientos es una tarea del Procesamiento del Lenguaje Natural que ha sido estudiada en diferentes dominios como el de películas, teléfonos móviles u hoteles. Sin embargo, otras áreas como el dominio médico no han sido exploradas todavía. En este trabajo presentamos un corpus de opiniones de pacientes formado por una parte en holandés (COPOD: Corpus of Patient Opinions in Dutch) y por otra parte en español (COPOS: Corpus of Patient Opinions in Spanish). Además, se han realizado diferentes experimentos en ambas lenguas utilizando un método supervisado (SVM), una aproximación basada en *cross-domain* y un método basado en diccionario. Los resultados obtenidos superan el método base en casi todos los casos e incluso los resultados de otros clasificadores de polaridad en el dominio del paciente. Con respecto al bilingüismo, los sistemas desarrollados para holandés y español proporcionan resultados similares para las medidas F1 y Accuracy.
Palabras clave: Clasificación de la polaridad, dominio médico, corpus de opiniones de pacientes, opener

1 Introduction

The examination of patients conducted by specialists when they suffer from some disease can mainly generate two types of information: i) medical reports with the personal and professional observations of physicians and ii) patient experiences. The experiences of these patients are sometimes published on the Internet generating a valuable information source that may contain not only facts but also opinions.

The field of study that analyses people's opinions, sentiments, evaluations, attitudes, and emotions from written language is known as Sentiment Analysis (SA) (Liu, 2012). In last years, the development and study of techniques for SA has been increased due to the vast amount of opinionated documents written on the Internet. Most of studies to date have focused on extracting opinions from user

ISSN 1135-5948

generated reviews in non-medical domains such as movies, phones or hotels. However, the study conducted by Fox and Duggan (2013) states that more than 85% of U.S. Internet users search online for health information. In addition, some platforms such as PatientsLikeMe[1] or Patient Opinion[2] expressing opinions related to health issues are becoming very popular. However, most of research on medical SA has been focused on English although interest in health-related information published in languages other than English is worldwide growing. For example, Van de Belt et al. (2013) present a study related to the preferences of the Dutch population revealing that 83% of people use the Internet as the main source for health-related information. In addition, 42.3% of Dutch population indicates that they sometimes search online for health-related information before visiting a physician. According to a study in 2015[3], 62% of the Spanish people consult the Internet to be informed about topics related to the health.

In this work we present a corpus of Dutch and Spanish patient reviews and apply sentiment analyses techniques in order to classify the reviews as positive or negative. Different methods are applied and compared: a supervised method (SVM), a cross-domain method (OpeNER) and a dictionary lookup method.

The rest of the paper is organized as follows: First we present an overview of related works, section 3 introduces the corpus, and section 4 describes the different methods. In section 5 the different experiments and their results are given. In section 6 the results are analyzed and discussed and we finish with section 7 that refers to conclusions and future works.

2 Related works

Despite of research in medical SA is scarce, there are some works dealing with opinions and sentiments in medical documents. A good review can be found in (Denecke and Deng, 2015). They consider 3 main areas of research in medical context according to the textual source: biomedical literature, clinical notes and medical web content. In this work we deal with patient opinions posted on the Web and we focus on polarity classification in order to identify whether the opinion expressed in a review is positive or negative. Thus, we are going to make a revision of the main studies on binary polarity classification in medical web content.

Qiu et al. (2011) present an interesting social science work in order to study how cancer survivors and caregivers benefit from participation in an online health community. The authors also apply Machine Learning (ML) techniques to analyse sentiment of 298 posts randomly selected from the Cancer Survivors Network[4]. AdaBoost with lexical and style features is the classifier that provides the best accuracy (79.2%).

Na et al. (2012) propose a rule-based linguistic approach for clause-level sentiment classification using existing resources such as UMLS, MPQA, SentiWordNet and Meta-Map. They test the approach over a set of 1,000 clauses of drug reviews manually labelled achieving an F1-measure of 0.70. Bobicev et al. (2012) build a corpus of tweets containing Personal Health Information and apply different ML algorithms to classify the sentiment expressed on the tweets with strong agreement between the annotators (669 tweets). The best result is obtained with a Naïve Bayes classifier (F1-measure=0.77).

Greaves et al. (2013) apply ML techniques to classify 6,412 online comments of patient experiences in hospitals of the English National Health Service. They also conduct binary classification experiments but using SA to capturing patient experience from texts. Their goal is to automatically predict whether a patient would recommend a hospital or not, whether the hospital was clean or not and whether the patient was treated with dignity or not. The best F1-measure values obtained in these experiments were 0.89 for hospital recommendation, 0.87 for cleanliness and 0.85 for respect. The algorithms that provided better results were multinomial Naïve Bayes and Bagging. Biyani et al. (2013) perform sentiment classification of user posts in an online health community (Cancer Survivor Network[5]) by exploiting domain-specific and general information features about sentiment expression and combining them in a semi-supervised setting using a co-training algorithm. The approach is tested on a set of

[1] https://www.patientslikeme.com

[2] https://www.patientopinion.org.uk/

[3] http://insights.doctoralia.es/informe-doctoralia-sobre-salud-e-internet-2015/

[4] http://csn.cancer.org

[5] http://www.csn.cancer.org

293 posts getting an F1-value of 0.84. Later, this work was extended in (Ofek et al., 2013). The authors show that classifiers trained using abstract features extracted from a dynamic sentiment lexicon outperform those trained using features extracted from a general sentiment lexicon. The number of features is reduced from thirteen to six and they obtain an F1-value of 0.81 with a Random Forest classifier.

Sharif et al. (2014) propose a representational richness framework that they evaluate on the AskaPatient dataset, a collection of 114,000 forum posts. The framework leverages novel feature representations that extract underlying sentiments in medical social media content. The feature set can be categorized into four categories: baseline features, semantic features, emotion related features and domain specific features. For evaluation, 24,000 posts are used for training and 90,000 for test with an SVM classifier getting an accuracy of 78.2%. Melzi et al. (2014) also apply an SVM classifier on a set of 3,000 sentences related to messages collected on the English-language Spine Health website. The best result is obtained with unigrams, bigrams, emotion words and patterns (F1 = 0,66).

All these studies are focused on English. However, recently, a corpus of 743 Spanish patient opinions extracted from the medical web Masquemedicos[6] has been presented (Plaza-del Arco et al., 2016). In order to demonstrate the usefulness of the resource, the authors conduct experiments using a general lexicon and a machine learning approach for the polarity classification of the reviews obtaining an F1 value of 0.72 and 0.71 respectively. This corpus is used in this paper in order to compare patient experiences in Dutch and Spanish, two languages of growing interest in health-related issues on the Web 2.0.

3 Resources

In this section, a detailed description is provided of the main resources employed in the experimental framework. We built a corpus of patient reviews that consists of a Dutch part (COPOD: Corpus of Patient Opinions in Dutch) and a Spanish part (COPOS : Corpus of Patient Opinions in Spanish).

COPOD, the Dutch part of the corpus, has been built by crawling the well-known medical forum Zorgkaart Nederland[7] on June 28, 2016. It is composed of 156,975 patient reviews about their experiences with physicians of 60 specialties. Each review contains information about the medical entity and the patient's opinion. In relation to the medical entity, the following elements have been extracted: the name, the profession, the specialty of the doctor and the city where the consultation was performed. With respect to the patient's opinion the following information has been included: the review text, the date, the disease treated, a rating for different aspects and an overall rating. The overall rating refers to a scale from 1 to 10 stars and corresponds to the average of the ratings of the different aspects (appointment, therapy, staff attention, information, listening, and accommodation). The number of reviews per rating is shown in Table 1. In addition, statistics of the corpus are shown in Table 3.

Rating (rt)	Reviews
$1 < rt <= 2$	1,781
$2 < rt <= 3$	2,302
$3 < rt <= 4$	3,690
$4 < rt <= 5$	4,290
$5 < rt <= 6$	3,459
$6 < rt <= 7$	2,870
$7 < rt <= 8$	12,465
$8 < rt <= 9$	49,577
$9 < rt <= 10$	76,541
Total	156,975

Table 1: COPOD - Reviews per rating

COPOS, the Spanish part of the corpus, is the first corpus of patient opinions in Spanish. It consists of 743 reviews about medical entities of 34 specialties that were extracted by crawling the medical forum Masquemedicos[8] on December 3, 2015. Each review contains information about the name and specialty of the medical entity, the city where the consultation was performed, the textual opinion, the date when the opinion was written and an overall evaluation with stars (from 0 to 5 stars). The number of reviews per rating can be seen in Table 2. Furthermore, some interesting features of the corpus are shown in Table 3.

[6] http://masquemedicos.com/

[7] https://www.zorgkaartnederland.nl/

[8] http://masquemedicos.com/

Rating (rt)	Reviews
0	3
1	88
2	18
3	35
4	51
5	548
Total	743

Table 2: COPOS - Reviews per rating

	COPOD	COPOS
#Sentences	534,317	2,009
#Words	7,341,779	32,365
Avg. sentences per review	3.4	2.7
Avg. words per sentence	13.7	16.1
Avg. words per review	46.7	43.6
#Adjectives	916,046 (12%)	3,002 (9%)
#Adverbs	540,355 (7%)	2,282 (7%)
#Nouns	1,173,732 (15%)	7,393 (22%)
#Verbs	1,111,525 (15%)	5,593 (17%)

Table 3: COPOD statistics

4 Methods

In this section we present the methods that have been used in the current research. We have carried out a set of experiments using an SVM classifier, a dictionary lookup method and OpeNER tool aiming at a binary classification of reviews.

4.1 Method I: SVM classifier

SVM classifier is based on the principle of Structural Risk Minimization of the computational learning theory (Vapnik, 2013). The theory is founded on the seeking of the hyper plane that maximizes the margin of separation between the objects belonging to two different classes. For the experimentation related to this study we applied 10-fold cross-validation and we used TF-IDF as weighting scheme and the libSVM implementation (Chang and Lin, 2011) with the following parameters: linear kernel, C=0.0 and epsilon=0.001.

4.2 Method II: Dictionary lookup method

The dictionary lookup method is a rule-based approach which starts from the review text and uses a sentiment lexicon to find positive and negative words in the review. The approach is a vote-algorithm: for each review the number of matched positive and negative words from the sentiment lexicon are counted. We then assign the majority polarity to the review. In the case of a tie positive polarity is assigned. The sentiment lexicon used to analyze the Dutch reviews is an automatically derived language sentiment lexicon obtained by WordNet propagation (cf. (Maks et al., 2014)). As this lexicon includes only lemma and part-of-speech we first lemmatized the text with the Dutch Alpino-parser[9]. The analysis of the Spanish reviews uses iSOL (Molina-González et al., 2013), that has been generated by translating into Spanish the Bing Liu English Lexicon. Both lexicons are general language lexicons and have not been adapted for the medical domain.

4.3 Method III: OPeNER cross-domain method

The cross-domain method makes use of the classifier that has been developed in the framework of the OpeNER project[10]. This project strived for the development of different opinion mining and sentiment analysis tools for several European languages including Dutch and Spanish. The set of tools includes classifiers that use Conditional Random Fields and are designed for finding opinion expressions in text. The tools have been trained on hotel reviews and our experiments aim at finding out how well these models can be applied in other domains. As the task at hand aims at classification at document level, instead of expression level, we calculate an overall opinion score by subtracting the number of negative expressions from the number of positive ones. If the result is smaller than zero the review is considered negative otherwise it is considered positive.

5 Experiments

As we have mentioned in Section 3, COPOS and COPOD were built in a similar way. Actually, both web sites (Zorgkaart Nederland

[9]http://www.let.rug.nl/ vannoord/alp/Alpino/
[10]http://www.opener-project.eu

	#Negative reviews	#Positive reviews	Total
COPOD	12,063	144,912	156,975
COPOS	109	634	743
COPOD-743	109	634	743

Table 4: Reviews per Corpus

and Masquemedicos) have analogous content about comments related to hospitals and doctors. Perhaps, the main difference lies in the rating scale since opinions in COPOD are ranking from 1 to 10 stars whereas in COPOS the scale is from 0 to 5 stars. On the other hand, COPOS is much smaller than the Dutch corpus. Due to this fact, we have also created a selection of COPOD that consists of 743 reviews with a similar distribution across rating categories as the Spanish corpus (COPOD-743). In this way, we can better compare the results using two comparable corpus in two different languages: Dutch and Spanish.

We have carried out a set of experiments using the SVM classifier, the dictionary lookup method and the OPeNER classifier. In our experiments we focus on a binary classification of the reviews. In order to select the positive and negative examples, we consider COPOD reviews with more than 5 stars as positive and the remaining reviews as negative. On the other hand, we consider positive reviews in COPOS if they have 3, 4 or 5 stars, and negative ones if their rating is 0, 1 or 2 stars. A summary of the number of reviews that composed each set is shown in Table 4.

In order to evaluate the different methods we calculated the usual measures per class: Precision (P) and Recall (R).

$$P_c = \frac{TP_c}{TP_c + FP_c} \quad (1)$$

$$R_c = \frac{TP_c}{TP_c + FN_c} \quad (2)$$

where TP (True Positives) are those assessments where the system and human experts agree on a label assignment, FP (False Positives) are those labels assigned by the system that do not agree with the expert assignment, FN (False Negatives) are those labels that the system failed to assign as they were given by the human expert, and TN (True Negatives) are those non-assigned labels that were also discarded by the expert. The Precision tells us how well the labels are assigned by our system (the fraction of assigned labels that are correct) whereas the Recall measures the fraction of the expert's labels found by the system. Finally, Precision and Recall are combined using the Macro-averaged F1 and Accuracy is measured in order to take into account all the correct results including TN (Sebastiani, 2002):

$$F1 - measure = \frac{1}{\mid c \mid} \sum_{i=1}^{|c|} \frac{2P_iR_i}{P_i + R_i} \quad (3)$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (4)$$

We created baseline measures assigning the most frequent class, i.e. positive to all reviews (cf. Table 5). Following, the experimentation results obtained with the different approaches over the three data sets are shown in Table 6, Table 7 and Table 8. It can be seen that libSVM provides the best accuracy results in all the datasets whereas the dictionary-based approach is around baseline and OpeNER is a bit below it.

6 Result analysis

According the corpus statistics, both COPOD and COPOS include a higher number of positive than negative opinions. It seems that patients are in general quite satisfied with the doctor's visit or they tend to write rather about their good experiences than about the bad ones.

One of the most salient results is that negative reviews have low performance across all methods when compared to positive reviews. With respect to the SVM method the reason can be found in the relative low number of negative training data but scarcity cannot affect the performance of the other methods. When comparing positive and negative reviews we found some characteristics that might explain it better. First of all, negative reviews are longer than positive ones: in COPOD the average length of positive reviews is 43.8 words

	COPOD	COPOS	COPOD-743
Majority baseline (positive class)	0.92	0.85	0.85

Table 5: Majority baselines

	COPOD	COPOS	COPOD-743
Precision negative class	0.76	0.90	0.91
Recall negative class	0.70	0.17	0.41
Precision positive class	0.97	0.88	0.89
Recall positive class	0.98	0.99	0.99
F1 measure	0.86	0.71	0.78
Accuracy	0.96	0.88	0.90

Table 6: Results for SVM

	COPOD	COPOS	COPOD-743
Precision negative class	0.52	0.60	0.66
Recall negative class	0.68	0.46	0.73
Precision positive class	0.97	0.91	0.94
Recall positive class	0.94	0.95	0.92
F1 measure	0.77	0.73	0.81
Accuracy	0.92	0.87	0.89

Table 7: Results for Dictionary based approach

	COPOD	COPOS	COPOD-743
Precision negative class	0.56	0.60	0.70
Recall negative class	0.10	0.08	0.12
Precision positive class	0.94	0.86	0.85
Recall positive class	0.98	0.99	0.99
F1 measure	0.56	0.53	0.56
Accuracy	0.91	0.86	0.85

Table 8: Results for OpeNER

whereas the average length of negative reviews is 71.8, and in COPOS these values are 38.2 and 74.5, respectively. A closer look at the texts reveals that negative reviews tend to describe events with a lot of detail and relatively often contain contextual opinions that require a wider context for correct interpretation. For example, one of the reviews contains the -in this case negative- expression *and after that an extra operation was needed*. It is only the broader context that explains that this extra operation was needed after an earlier surgery that went wrong for unnecessary reasons. This kind of expressions are hard to automatically identify and classify. Secondly, we noted that relatively many negative reviews are in the middle rating categories (3,4 and 5 stars for COPOD and 2 and 3 for COPOS) whereas most positive reviews are in the extreme rating categories (9 and 10 stars for COPOD and 5 for COPOS). That may also explain low performance on the negative class as earlier research (cf. (Maks and Vossen, 2013)) already showed that reviews of the middle rating categories are hard to classify because they often include a mixture of positive and negative opinions. For example, in the specific case of COPOD we have realized that comments rating with 6, 7 or even 8 stars could be considered semantically negative when you read the textual information, although we have taken as negative reviews until 5 stars. Thus, a better partition of the corpus considering for instance negative re-

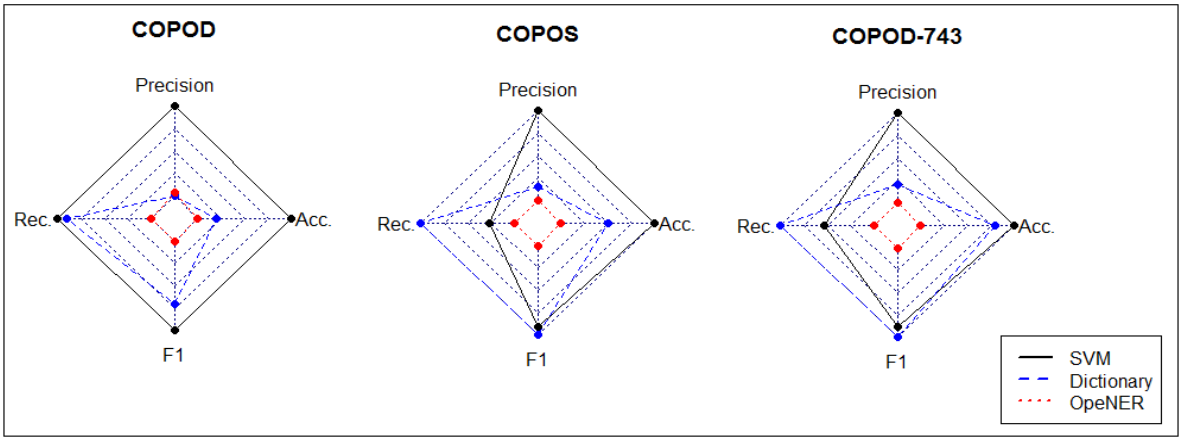

Figure 1: Analysis of classification methods using radar charts (Larger area implies better overall performance of the method)

views between 0 and 7 stars and positive ones whose between 8 and 10 could improve the final results and help to correctly classify negative examples.

On the other hand, it is interesting to note that the behavior of the systems are very similar for both languages, always presenting a better performance for positive class than for negative one. As expected the dictionary lookup method and the OpeNER method have lower overall performance than SVM (Figure 1) as both are methods not adapted for the medical domain. Another difference is that SVM's bag of words approach works at document level whereas the other 2 methods work at expression level identifying and classifying each separate opinion expression. Although expression-level classification may not be the best approach for the current task we think that is needed for more fine-grained tasks such as, for example, aspect-based sentiment analysis.

7 Conclusion

In this paper we have presented a corpus with patient reviews written in Dutch and Spanish. We have conducted different experiments using a supervised method, a cross-domain method and a dictionary lookup method.

Research in medical domain for SA is very scarce and this paper present a background with the main works of the area. On the other hand, most of research is focused on English although interest in subjective medical information is growing in other languages. For this reason, we have centered our work on Dutch and Spanish and we have presented several approaches to tackle the problem. The results show low differences between languages and, although the SVM method has a better performance, the dictionary approach also reaches good accuracy. Perhaps the worst result is obtained with the cross-domain approach, but we must take into account that the OpeNER tool has been trained over the tourism domain and it has been directly applied to the medical domain.

Finally, we consider this paper as a preliminary research and our future work will be focused on other issues related to SA for health such as the study of aspect-based SA in medical domain using the generated corpus or the generation of resources adapted to the medical domain.

Acknowledgments

This work has been partially supported by a grant from the Ministerio de Educación Cultura y Deporte (MECD - scholarship FPU014/00983), Fondo Europeo de Desarrollo Regional (FEDER) and REDES project (TIN2015-65136-C2-1-R) from the Spanish Government.

References

Biyani, P., C. Caragea, P. Mitra, C. Zhou, J. Yen, G. E. Greer, and K. Portier. 2013. Co-training over domain-independent and domain-dependent features for sentiment analysis of an online cancer support community. In *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), 2013*, pages 413–417. IEEE.

Bobicev, V., M. Sokolova, Y. Jafer, and

D. Schramm. 2012. Learning sentiments from tweets with personal health information. In *Canadian Conference on Artificial Intelligence*, pages 37–48. Springer.

Chang, C.-C. and C.-J. Lin. 2011. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27.

Denecke, K. and Y. Deng. 2015. Sentiment analysis in medical settings: New opportunities and challenges. *Artificial intelligence in medicine*, 64(1):17–27.

Fox, S. and M. Duggan. 2013. Health online 2013. *Health*, pages 1–55.

Greaves, F., D. Ramirez-Cano, C. Millett, A. Darzi, and L. Donaldson. 2013. Use of sentiment analysis for capturing patient experience from free-text comments posted online. *Journal of medical Internet research*, 15(11):e239.

Liu, B. 2012. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167.

Maks, I., R. Izquierdo, F. Frontini, R. Agerri, and P. Vossen. 2014. Generating Polarity Lexicons with Wordnet propagation in five languages. In *Proceedings of LREC2014*, Reykjavik.

Maks, I. and P. Vossen. 2013. Sentiment Analysis of Reviews: Should we analyze writer intentions or reader perceptions? In *Proceedings of RANLP 2003*, pages 415–419, Hissar, Bulgaria.

Melzi, S., A. Abdaoui, J. Azé, S. Bringay, P. Poncelet, and F. Galtier. 2014. Patient's rationale: Patient Knowledge retrieval from health forums. In *eTELEMED: eHealth, Telemedicine, and Social Medicine*.

Molina-González, M. D., E. Martínez-Cámara, M.-T. Martín-Valdivia, and J. M. Perea-Ortega. 2013. Semantic orientation for polarity classification in spanish reviews. *Expert Systems with Applications*, 40(18):7250–7257.

Na, J.-C., W. Y. M. Kyaing, C. S. Khoo, S. Foo, Y.-K. Chang, and Y.-L. Theng. 2012. Sentiment classification of drug reviews using a rule-based linguistic approach. In *International Conference on Asian Digital Libraries*, pages 189–198. Springer.

Ofek, N., C. Caragea, L. Rokach, P. Biyani, P. Mitra, J. Yen, K. Portier, and G. Greer. 2013. Improving sentiment analysis in an online cancer survivor community using dynamic sentiment lexicon. In *Social Intelligence and Technology (SOCIETY), 2013 International Conference on*, pages 109–113. IEEE.

Plaza-del Arco, F. M., M. T. Martín-Valdivia, S. M. Jiménez-Zafra, M. D. Molina-González, and E. Martínez-Cámara. 2016. COPOS: Corpus Of Patient Opinions in Spanish. Application of Sentiment Analysis Techniques. *Procesamiento del Lenguaje Natural*, 57:83–90.

Qiu, B., K. Zhao, P. Mitra, D. Wu, C. Caragea, J. Yen, G. E. Greer, and K. Portier. 2011. Get online support, feel better–sentiment analysis and dynamics in an online cancer survivor community. In *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on*, pages 274–281. IEEE.

Sebastiani, F. 2002. Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1):1–47, March.

Sharif, H., F. Zaffar, A. Abbasi, and D. Zimbra. 2014. Detecting adverse drug reactions using a sentiment classification framework.

Van de Belt, T. H., L. J. Engelen, S. A. Berben, S. Teerenstra, M. Samsom, and L. Schoonhoven. 2013. Internet and social media for health-related information and communication in health care: preferences of the Dutch general population. *Journal of medical Internet research*, 15(10):e220.

Vapnik, V. 2013. *The nature of statistical learning theory*. Springer Science & Business Media.