Computational Linguistics in the Netherlands Journal 3 (2013) 135-147 Submitted 05/2013; Published 12/2013

Sentiment classification for early detection of health alerts in the chemical textile domain

Javi Fernández* JAVIFM@DLSI.UA.ES
Carolina Prieto† CPRIETO@AITEX.ES
Elena Lloret* ELLORET@DLSI.UA.ES
José M. Gómez* JMGOMEZ@DLSI.UA.ES
Patricio Martínez-Barco* PATRICIO@DLSI.UA.ES
Manuel Palomar* MPALOMAR@DLSI.UA.ES

*Department of Software and Computing Systems, University of Alicante
†AITEX, Textile Industry Research Association

Abstract

In the *chemical textile* domain experts have to analyse chemical components and substances that might be harmful for their usage in clothing and textiles. Part of this analysis is performed searching opinions and reports people have expressed concerning these products in the Social Web. However, this type of information on the Internet is not as frequent for this domain as for others, so its detection and classification is difficult and time-consuming. Consequently, problems associated to the use of chemical substances in textiles may not be detected early enough, and could lead to health problems, such as allergies or burns. In this paper, we propose a framework able to detect, retrieve, and classify subjective sentences related to the chemical textile domain, that could be integrated into a wider *health surveillance* system. We also describe the creation of several datasets with opinions from this domain, the experiments performed using machine learning techniques and different lexical resources such as WordNet, and the evaluation focusing on the *sentiment classification*, and *complaint detection* (i.e., *negativity*). Despite the challenges involved in this domain, our approach obtains promising results with an F-score of 65% for polarity classification and 82% for complaint detection.

1. Introduction

Currently, the Web provides large amounts of information that users find interesting for general use. The creation of Web 2.0 (i.e., Social Web) has allowed users to have an active participation through their comments and opinions, stated about a wide range of topics and/or services (e.g., products, restaurants, hotels, etc.). Therefore, besides objective and factual information, we can also find a lot of subjective information, which is expressed through new textual genres, such as blogs, forums, reviews, social networks and microblogs, among others. This subjective information has a great value for both general and expert users, but it is difficult to exploit it accordingly. In some cases the amount of subjective information can be too hard to find but at the same time it must be detected as soon as possible. This is the case of the *chemical textile* domain.

In this domain, experts have to analyse chemical components and substances that might be harmful for their usage in clothing and textiles. They analyse the information available on the Social Web, searching for comments, reports, opinions, etc. that people have expressed concerning specific products or components. Normally, if there is something wrong with any textile or clothing, the number of complaints on the Web increases. In some cases, these complaints can be considered as alerts, that will be studied by some specialised quality committee, and may result in a more deep analysis about the product or component that have been put to complaints.

An example of such a case happened between 2006 and 2008 with the component *Dimethylfumarate* (DMFu)[1][2]. This is a biocide product used in the process of transport. It was pulverised on all type of products transported from Asia to Europe, to avoid humidity in the products while they were transported. DMFu was not known in Europe and was neither registered nor prohibited in any European legislation. Users started to write lots of negative comments in blogs and forums, concerning the reactions it produced (itching, irritation, redness, burns, etc.). In the next months, all these complaints reached the Ministry of Health, who raised the alarm by establishing this component as dangerous and started an investigation. After the study of this product and the appropriate tests, the use of DMFu was finally prohibited in Europe. If experts would have had access to all this information in advance through automatic tools, they could have solved this issue much before, avoiding serious health problems in hundreds of victims (Ferguson et al. 2005, Longini et al. 2005).

Most of the current *health surveillance systems* can help in this purpose, automating much of the work experts carry out. However, for this kind of systems, this domain is very challenging, because the amount of information on the Internet is very small compared to other domains (e.g. influenza pandemics detection). In addition, they do not differentiate between a document talking about a product and a document complaining about it. Therefore, an additional *sentiment analysis* process is required for detecting negative comments (Chanlekha et al. 2010, Chew and Eysenbach 2010). In this paper, we propose a framework able to detect, retrieve, and classify subjective sentences related to the chemical textile domain, that could be integrated into a wider *health surveillance* system.

Furthermore, for a sentiment classification system, this domain is also very challenging compared to the traditional ones (e.g., movies, technology, politics, etc.). The amount of information available regarding this domain is not as large as in other ones. In addition, the number of complaints is usually much bigger than the number of positive opinions. This causes a lack of balance that current systems find very difficult to deal with. Therefore, through our research, we want to analyse to what extent a sentiment classification system could be beneficial and useful for experts working in the chemical textile domain, studying their potentials and limitations. The experiments performed are focused on the task of sentiment classification.

The remainder of the paper is structured as follows. In Section 2, we briefly describe the related work in these areas (sentiment analysis and health surveillance). In Section 3, we describe the framework proposed, the terminology employed, as well as the the tools and resources used in the implementation. Section 4 explains the corpus we collected and used in this study and its annotation process. The experiments performed and their evaluation and discussion are provided in Section 5. Finally, Section 6 concludes the paper, and outlines the future work.

2. Related work

2.1 Sentiment analysis

Sentiment analysis (SA) is the task of identifying the opinions expressed in text and classifying texts accordingly (Dadvar et al. 2011). In this task two main approaches can be followed (Annett and Kondrak 2008, Liu 2010, Taboada et al. 2011): *lexical* approaches (unsupervised SA) and *machine learning* approaches (supervised SA). Lexical approaches focus on building dictionaries and lexicons of labelled words. This labelling indicates not only if a word represents a positive or a negative opinion, but also its intensity. If a word is found in a text, its polarity value is added to the total polarity score of the text. If this total score is positive, then that text is classified as positive, otherwise it is classified as negative. These dictionaries can be created manually (Stone et al. 1966) or automatically, using seed words to expand the list of words (Turney 2002). Examples of lexicons

1. http://europa.eu/rapid/press-release_IP-09-190_en.htm
2. http://www.dailymail.co.uk/femail/article-1028097/This-baby-burned-red-raw-sofa-giving-toxic-fumes-As-investigation-reveals-hundreds-victims.html

are *WordNet Affect* (Strapparava and Valitutti 2004), *SentiWordNet* (Esuli and Sebastiani 2006), *MicroWNOP* (Cerini et al. 2007) or *JRC Tonality* (Balahur et al. 2009). However, it is very difficult to collect and maintain a universal sentiment lexicon for all application domains because different words may be used in different domains (Qiu et al. 2009) and some words are domain dependent (Turney 2002).

The other approach uses machine learning techniques. These techniques imply the creation of a corpus containing a set of classified texts for training a classifier, which can then be applied to classify a set of unclassified texts. The majority of the researches employ *Support Vector Machines* (Mullen and Collier 2004, Prabowo and Thelwall 2009, Wilson et al. 2005) or *Naïve Bayes* (Pang and Lee 2004, Wiebe and Riloff 2005, Tan et al. 2009) classifiers because they usually obtain the best results. In this approach, texts are represented as vectors of features, and depending on the features used the system can reach better results (bag-of-words and lexeme-based features are the more commonly used for these tasks (Pang and Lee 2008)). Moreover, there also has been work using the part of speech (POS) to use only ones with a specific role (adjectives are the most common features (Pang and Lee 2008)). Lexical resources such as *WordNet* can also be exploited to take advantage of the semantic information they provide (Balamurali et al. 2011). These classifiers perform very well in the domain that they are trained on, but their performance drops when the same classifier is used in a different domain (Pang and Lee 2008, Tan et al. 2009).

In both approaches, the problem of *domain dependence* is common. When the lexicons and classifiers generated are used in a domain different from the one they were built for, they usually perform worse (Turney 2002, Pang and Lee 2008, Qiu et al. 2009, Tan et al. 2009). Creating a domain-specific lexicon or classifier means making a manual effort. Although some studies try to overcome this problem by generating the lexicons *automatically* (Turney 2002), learning from *unannotated* texts (Wiebe and Riloff 2005) or using *hybrid* approaches (Andreevskaia and Bergler 2008, Bollen et al. 2009, Zhang and Ye 2008), a minimal intervention from experts in the domain is needed. In this study we use the machine learning approach due to the good results obtained in previous works (Boldrini et al. 2009, Fernández et al. 2011).

2.2 Public Health Surveillance

Public Health Surveillance (PHS) is defined as the ongoing systematic collection, analysis, and interpretation of data essential to the planning, implementation, and evaluation of public health practice, closely integrated with the timely dissemination of these data to those who need to know (Thacker and Berkelman 1988). There has been a growing interest in monitoring disease outbreaks using the Internet by mining newspapers (Linge et al. 2009), health-related websites (Brownstein et al. 2008), social networks and microblogs (Culotta 2010, Chew and Eysenbach 2010) or search engines (Ginsberg et al. 2008, Eysenbach 2006). While many rely on keyword matching or document classification, some apply more complex linguistic analysis such as named-entity recognition and topic modeling (Collier et al. 2008, Brownstein et al. 2008) and only a few use sentiment analysis (Culotta 2010). Moreover, Chanlekha et al. (2010) and Chew and Eysenbach (2010) emphasise the need of using sentiment analysis to improve current health surveillance systems.

3. Description of the system

In this section we describe our approach to detect, retrieve, and classify subjective sentences related to the chemical textile domain, integrated into a *health surveillance* system. In Figure 1 we can see the general structure of our approach. At the *Term Selection* stage the experts select a list of terms to monitor. These terms are used as queries for the *Document Retrieval* system, obtaining a list of relevant documents. They are properly processed using *natural language processing* tools to obtain the sentences mentioning any of the selected terms, at the *Document Processing* stage. Using a *classifier*, the *Sentiment Classification* system detects the sentences which have opinionated

information. Finally, a report is generated summarising all this data for the experts, who will decide if the report is indicating a health alarm. In order to create the classifier, we extract a subset of the sentences and *manually annotate* them to create a corpus. They will be used by the *Machine Learning* algorithm to create the classifier. Following this, the description is explained in depth:

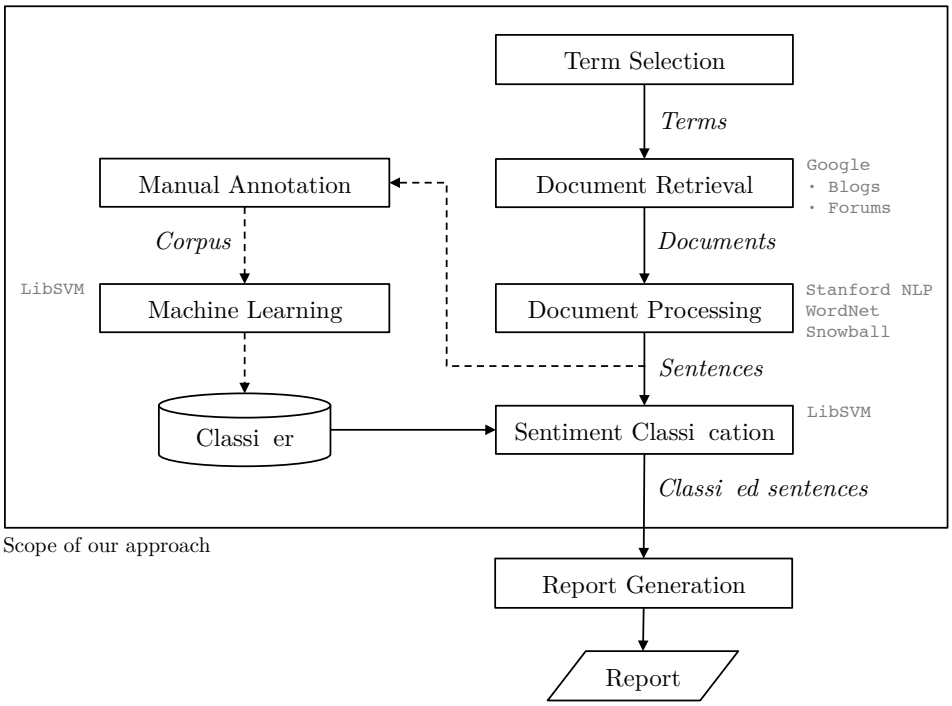

Figure 1: Structure of our approach using sentiment classification

1. *Term Selection.* In our approach, the expert decides what terms are the most relevant ones to watch within the topic they are interested in. In the chemical textile domain, the directives and legislation include the most important chemical components and substances in this domain, such as Reach[3], as well as different reports of testing in ecological certification Oeko-tex Standard 100[4]. For this study we selected 50 of those components as terms, shown in Table 1.

2. *Document Retrieval.* The system searches in the Social Web the terms selected by the user obtaining the most relevant documents for each one. Not only the relevance is important at this stage but also the date of publication of those documents. As the system must provide updated information to the user, the more recent the document is, the more important we consider it. Any information retrieval system or search engine can be used at this stage, but taking into account that the documents retrieved *must be as recent as possible.*

   In our implementation, we chose the English version of *Google Search*[5] as document retrieval system, because the documents it returns are usually updated and it has the capability to sort the results by date, so we can retrieve the most recent ones. In addition, this search engine allows us to filter the documents retrieved by the textual genre they come from (e.g., news, blogs, books, forums, etc.). In this case, we select only those which belong to *blogs* or *forums*, because it is more probable to find opinions within these genres. In a future version

3. http://www.reachinnova.com/
4. http://www.oeko-tex.com/
5. http://www.google.com/

dimetylfumarate · formaldehyde · heavy metals · antimony
arsenic · lead · cadmium · chromium · cobalt · copper · nickel
mercury · pesticides · permethrine · hexachlorobenzene
captafol · chlorinated phenols · phthalates · fluorene
di-iso-nonylphtalate · di-n-octylphtalate · naphtalene
di-(2-ethylhexyl)-phtalate · organic tin compounds
tributyltin · triphenyltin · dibutyltin · dioctyltin
chlorinated benzenes and toluenes · decabromodiphenylether
polycyclic aromatic hydrocarbons · octabromodiphenylether
pyrene · phenanthrene · anthracene · fluoranthene · arylamines
flame retardants · hexabromocyclododecano · dimethylacetamide
polybrominated biphenyles · orthophenylphenol · dimethylformamide
chlorinated paraffins · nonylphenolethoxilates
octylphenolethoxylates · perfluorooctane sulfonates
dyestuffs carcinogenic · dyestuffs allergenic

Table 1: List of components used as terms for the analysis

of this system we will increase the coverage by including review sites and social networks like *Twitter*[6] (searching in the public timeline) and *Facebook*[7] (following the public updates from specialised pages, groups and individuals), in addition to other Social Web textual genres. Using the search engine we perform automatically a query for each selected term (Table 1), obtaining a list of documents for each one.

3. *Document Processing.* At this stage we have a list of documents from blogs and forums, so we know their format is HTML[8] and we can easily obtain their text by removing tags and scripts. Next, documents are split into sentences using the *Stanford CoreNLP* library[9]. Finally, sentences not containing the selected terms are removed automatically. Our approach assumes that sentences containing one of the pre-selected terms are relevant, regardless of whether the target of the opinion is actually one of them. This approach can entail some errors, therefore as future work we will incorporate target detection techniques to obtain more accurate results.

4. *Manual Annotation.* As the system uses the machine learning approach for the polarity classification, we must build a training corpus. A subset of the sentences retrieved were manually classified as *positive*, *negative* or *neutral*. We describe in depth the annotation process and the corpus creation in Section 4.

5. *Machine Learning.* We use the manually annotated sentences to generate a classifier using machine learning techniques. We consider polarity classification as a text classification task (Sebastiani 2002, Pang and Lee 2008), where the polarities annotated are used as categories and the terms are used as features. Different types of terms are extracted for each sentence: *words*, *stems*, *lemmas*, *word n-grams*, *stem n-grams*, *lemma n-grams* and *synsets*. This extraction is also performed using the *Stanford CoreNLP* library. *Words* and *lemmas* are directly obtained from this parser (Toutanova et al. 2003). Using the *Snowball* implementation[10] of the *Porter*

6. `http://www.twitter.com/`
7. `http://www.facebook.com/`
8. `http://www.w3.org/html/`
9. `http://nlp.stanford.edu/software/corenlp.shtml`
10. `http://snowball.tartarus.org`

Stemmer (Porter et al. 1980) algorithm, we obtain the *stems*. Using the *lemmas* and the *part-of-speech* obtained from the parser, we obain the term *synsets* from *WordNet 3.0* (Fellbaum 1998) (synsets are sets of synonyms identified by an unique number). The *word n-grams*, *stem n-grams*, *lemma n-grams* (unigrams, bigrams and trigrams) are obtained by generating new terms of consecutive *words*, *stems*, and *lemmas* respectively. Finally, all these elements are weighted using *normalised tf-idf* (Sebastiani 2002) and used as features for the machine learning algorithm. As supervised machine learning method we use *Support Vector Machines* (SVM) due to its good performance in text categorisation (Sebastiani 2002) and previous works in sentiment analysis (Pang and Lee 2004, Mullen and Collier 2004, Wilson et al. 2005, Prabowo and Thelwall 2009, Boldrini et al. 2009, Fernández et al. 2011). More specifically, we use the *Weka*[11] *LibSVM*[12] implementation (Hall et al. 2009, Chang and Lin 2011) with the *Radial basis function* kernel and the default parameters.

6. *Sentiment Classification.* The system classifies the polarity of each sentence automatically using the generated classifier and, finally, it returns a list of classified sentences.

7. *Report Generation.* The classified sentences are processed to give a detailed summary of the opinions about the selected terms. In this study we are focused on sentiment classification so did not implement this stage, but we will in future work.

4. Corpus

As we did not find any sentiment corpus in the chemical textile domain, it was necessary to create one[13], using the sentences obtained at the document retrieval stage. We made a team composed by 16 people, familiarised with *natural language processing* (NLP) and the chemical textile domain. We gave them a period of *two weeks* to annotate as many sentences as possible. They had to classify the *polarity* of each sentence, depending on whether it contained a *positive* or *negative* opinion, or it was a *neutral* sentence. To help in this purpose, an on-line annotation tool was developed. It made the annotation process easier allowing annotators to focus on one sentence at a time, and assured us that every sentence was annotated by three different people. In Figure 2 an example of the web interface is shown.

# of sentences retrieved	476,975
# of sentences containing pre-selected terms	2,253
# of sentences annotated by 3 people	870
# of sentences with a minimum agreement of 2	671
# of sentences with a minimum agreement of 3	285

Table 2: Annotation Statistics

Once the period expired, we closed the annotation process and created the corpora for this study. In Figure 2 we can see the annotation statistics. We obtained a moderate agreement with a kappa value of 0.45 (Fleiss 1971, Landis and Koch 1977). Two datasets were built, in order to check if a higher annotation quality (annotations with a higher agreement) has a noticeable influence in the system performance. The first one contains those sentences with an agreement of at least two different people (corpus *Polarity-A2*), and the second one contains the sentences where all 3 annotators agreed (corpus *Polarity-A3*). It is important to remark that the Polarity-A3 dataset is a subset of the Polarity-A2 dataset. The rest of the sentences were rejected because there was no agreement in their classification. In Table 3 we can see the statistics for these corpora.

11. `http://www.cs.waikato.ac.nz/ml/weka/`
12. `http://www.csie.ntu.edu.tw/~cjlin/libsvm/`
13. Available on request from authors

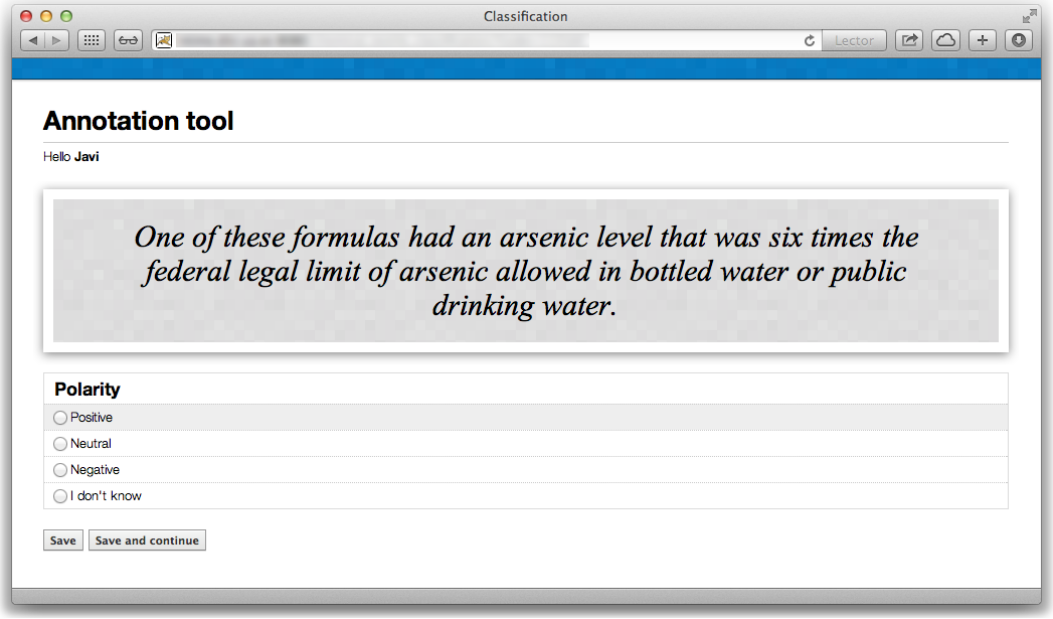

Figure 2: Annotation tool developed for this study. Only one sentence can be annotated at a time.

Corpus	Positive	Neutral	Negative	Total
Polarity-A2	141	294	236	671
Polarity-A3	55	111	119	285

Table 3: Size of the polarity corpora by category and agreement

However, in the chemical textile domain, the experts are more interested in detecting negative opinions rather than the positive ones. This is the reason we derived two additional corpora from the Polarity datasets, joining the positive and neutral categories into one category, called *not negative*. Hence the sentences here are classified only in two categories: *negative* and *not negative*. As before, it is also divided into two corpora, depending on the agreement: *Negativity-A2* and *Negativity-A3*. In Table 4 we can see the statistics for these corpora.

Corpus	Negative	Not negative	Total
Negativity-A2	236	435	671
Negativity-A3	119	166	285

Table 4: Size of the negativity corpora by category and agreement

5. Evaluation

We performed a series of experiments to evaluate both polarity and negativity classification using different types of terms (words, lemmas, synsets, etc.) to check which ones are the best for this task. All these experiments share the following common parameters:

- The evaluation is performed using *10-fold cross validation* because of the small size of the corpus.

- The measures used are the traditional ones: *precision* and *recall*. We do not use *accuracy* because it is not a good measure for text categorisation when using an imbalanced corpus (Yang and Liu 1999). Instead, we also use the *F-score* (F1) because it represents a balance between precision and recall.
- As *baseline* for all the experiments we use a classifier that always assigns the most frequent category to every text in the test set of each fold.

5.1 Evaluation of Polarity

In this first set of experiments we used the *Polarity* corpora. As previously mentioned, these datasets contain sentences classified as *positive*, *neutral* or *negative*. We made an experiment for each type of feature (words, stems, lemmas, etc.). In the case of the synsets, *WordNet* does not provide a single synset per word but a list of all possible synsets ranked by frequency (the first sense is more likely than the second, the second is more likely than the third, etc.). As we do not use any *word sense disambiguation* (WSD) tool we decided to adopt two approaches: using all the synsets from each term (experiment *All synsets*) and using only the first synset (experiment *First synset*). In Table 5 we can see the results obtained.

Corpus	Tokens	Precision	Recall	F1
Polarity-A2	Baseline	.1920	.4382	.2670
	Words	.6035	.5917	.5672
	Stems	.6121	.5976	.5693
	Lemmas	.6145	**.6006**	**.5707**
	Word n-grams	.6460	.5797	.5265
	Stem n-grams	.6305	.5797	.5287
	Lemma n-grams	**.6735**	.5857	.5318
	First synset	.6267	**.6006**	.5682
	All synsets	.5457	.5648	.5282
Polarity-A3	Baseline	.1743	.4175	.2460
	Words	.6346	.6211	.5936
	Stems	.6866	.6632	.6459
	Lemmas	.6656	.6386	.6265
	Word n-grams	.6753	.5965	.5354
	Stem n-grams	.6651	.5895	.5272
	Lemma n-grams	.6594	.5825	.5306
	First synset	**.7064**	**.6667**	**.6501**
	All synsets	.6132	.6000	.5780

Table 5: Results for Polarity

These data must be interpreted with caution due to the small size of the datasets, as the findings might not be transferable to other domains. The results obtained with the *Polarity-A3* corpus are slightly better than the ones obtained with the *Polarity-A2* corpus, despite the first one being much smaller. A possible explanation for this might be that the agreement is higher in the A3 corpus, so it can thus be suggested that the quality of the annotation is important for the training process. It is likely therefore that a higher agreement implies the sentences are clearer or they have terms that emphasise their polarity.

Although using only words significantly outperformed the baseline results, the best results are obtained using lexical resources. The best results are obtained using the *Polarity-A3* corpus and using the most frequent synsets as tokens, reaching a precision of 70% and a F-score of 65%. Using senses instead of words increases the recall because a single sense groups several words and, therefore,

the classifier can recognise more words. Using all the senses does not perform well because it adds too much noise to the learning algorithm.

When using n-grams the precision reaches a good level. This behaviour may be explained by the fact that n-grams can coincide with domain-specific expressions, so the precision is high. Nevertheless, the recall decreases significantly, because in a small dataset the probability of an n-gram to be repeated is very low.

5.2 Evaluation of Negativity

In this second set of experiments we used the Negativity corpora. These corpora contain sentences classified as *negative* or *not negative*. The experiments performed are the same as the ones performed in the previous section. In Table 6 we can see the results obtained.

Agreement	Tokens	Precision	Recall	F1
Negativity-A2	Baseline	.4203	.6483	.5100
	Words	.7768	.7779	.7658
	Stems	**.7857**	.7869	**.7765**
	Lemmas	.7744	.7765	.7649
	Word n-grams	.7705	.7601	.7354
	Stem n-grams	.7765	**.7871**	.7562
	Lemma n-grams	.7711	.7630	.7407
	First synset	.7477	.7541	.7443
	All synsets	.7201	.7288	.7125
Negativity-A3	Baseline	.3393	.5825	.4288
	Words	.7564	.7579	.7550
	Stems	.7672	.7684	.7659
	Lemmas	.7598	.7614	.7592
	Word n-grams	.7963	.7860	.7777
	Stem n-grams	.7945	.7860	.7784
	Lemma n-grams	.8021	.7965	.7908
	First synset	**.8282**	**.8246**	**.8211**
	All synsets	.7330	.7333	.7267

Table 6: Results for Negativity

Again, the results obtained with the *Negativity-A3* corpus are slightly higher than the ones obtained with the *Negativity-A2* corpus. All the experiments outperformed the baseline results and the best results are obtained using lexical resources, more specifically using the *Negativity-A3* corpus and the most frequent synsets as tokens, reaching a precision of almost 83% and a F-score of 82%.

All the experiments using the Negative corpora achieve higher scores than the ones using the Polarity ones. In the case of the Negativity datasets, the number of categories is lower, so it is more probable to assign the correct category to each sentence. We observe this fact looking at the baselines for each dataset: the ones in the Negativity corpora are nearly twice the ones in the Polarity corpora. In addition, the Negativity datasets are less imbalanced than the Polarity ones, so the performance of the machine learning algorithm in the Polarity corpora is likely to decrease, as demonstrated by Kang and Cho (2006) and Borrajo et al. (2011).

As the main conclusion, we can deduce that it is very important to distinguish the requirements and peculiarities of the domain when designing the system. In the case of the chemical textile domain, the most important issue is to discern what opinions are negative and what are not, resulting in a system with less categories and, therefore, improve its performance.

6. Conclusions

In this paper, we proposed a framework able to detect, retrieve, and classify subjective sentences related to the chemical textile domain, that could be integrated into a wider *health surveillance* system. We built several sentiment corpora with documents from this domain retrieved from the Social Web and performed a series of experiments using machine learning techniques with different tools and lexical resources. Despite the challenges this domain has, our approach obtained promising results, with an F-score of 65% for sentiment polarity and 82% for negativity as the best ones. The application of sentiment analysis to the chemical textile domain is very useful and innovative, encouraging us to continue with the research and development of our system. We can extract some main conclusions from this study:

- The results of this research support the idea that the quality of the annotation is important for the training process. Using the texts with higher agreement suggests the sentences are clearer or they have terms that emphasise their polarity.

- This study has shown that, in this domain, the lexical resources are very useful for the sentiment classification task. The best results were obtained using the most frequent synset from *WordNet*.

- It is important when designing the system to distinguish the requirements of the user and the domain. The number of categories and the distribution of texts within each category can improve considerably the performance of the system. In the case of the chemical textile domain, studying the negativity instead of the polarity results in an improvement of the sentiment classification process and, therefore, in the general performance of the system.

As future work we propose the following tasks:

- Increase the number of annotated sentences to check if the results obtained are not dependent on the current size of the datasets.

- Employ machine learning algorithms different from SVM, especially algorithms such as *Hidden Markov Models*, which have into account the sequentiality of the terms in the text.

- Apply advanced target detection techniques, such as *named entity recognition* and *semantic roles*, to check if the pre-selected terms are actually the target of the opinions and complaints.

- Add *parsing information* to the classifier to handle negation and modality.

- Use a domain specific ontology to avoid the ambiguity of the terms introduced by the user and retrieve documents as relevant as possible.

- Include review sites and social networks, in addition to other Social Web textual genres.

- Implement and evaluate the system in a different non traditional domain (such as medicine).

Acknowledgements

We would like to express our gratitude for the financial support given by the Department of Software and Computer Systems at the University of Alicante, the Spanish Ministry of Economy and Competitivity (Spanish Government) by the project grants TEXT- MESS 2.0 (TIN2009-13391-C04-01), LEGOLANG (TIN2012-31224), and the Valencian Government (grant no. PROMETEO/2009/119).

References

Andreevskaia, A. and S. Bergler (2008), When specialists and generalists work together: Overcoming domain dependence in sentiment tagging, *Proceedings of ACL-08: HLT* pp. 290–298.

Annett, M. and G. Kondrak (2008), A comparison of sentiment analysis techniques: Polarizing movie blogs, *Advances in Artificial Intelligence* pp. 25–35, Springer.

Balahur, A., R. Steinberger, E. Goot, B. Pouliquen, and M. Kabadjov (2009), Opinion mining on newspaper quotations, *Web Intelligence and Intelligent Agent Technologies, 2009. WI-IAT'09. IEEE/WIC/ACM International Joint Conferences on*, Vol. 3, IEEE, pp. 523–526.

Balamurali, A.R., A. Joshi, and P. Bhattacharyya (2011), Robust sense-based sentiment classification, *ACL HLT 2011* p. 132.

Boldrini, E., J. Fernández Martínez, J.M. Gómez Soriano, P. Martínez-Barco, et al. (2009), Machine learning techniques for automatic opinion detection in non-traditional textual genres, WOMSA.

Bollen, J., A. Pepe, and H. Mao (2009), Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena, *Proc. of WWW 2009 Conference*.

Borrajo, L., R. Romero, E.L. Iglesias, and C.M. Redondo Marey (2011), Improving imbalanced scientific text classification using sampling strategies and dictionaries, *Journal of Integrative Bioinformatics* **8** (3), pp. 176.

Brownstein, J.S., C.C. Freifeld, B.Y. Reis, and K.D. Mandl (2008), Surveillance Sans Frontiéres: Internet-based emerging infectious disease intelligence and the HealthMap project, *PLoS medicine* **5** (7), pp. e151, Public Library of Science.

Cerini, S., V. Compagnoni, A. Demontis, M. Formentelli, and G. Gandini (2007), Micro-WNOp: A gold standard for the evaluation of automatically compiled lexical resources for opinion mining, *Language Resources and Linguistic Theory: Typology, Second Language Acquisition, English Linguistics* pp. 200–210.

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), pp. 27, ACM.

Chanlekha, H., A. Kawazoe, and N. Collier (2010), A framework for enhancing spatial and temporal granularity in report-based health surveillance systems, *BMC Medical Informatics and Decision Making* **10** (1), pp. 1, BioMed Central Ltd.

Chew, C. and G. Eysenbach (2010), Pandemics in the age of Twitter: content analysis of tweets during the 2009 h1n1 outbreak, *PLoS One* **5** (11), pp. e14118, Public Library of Science.

Collier, N., S. Doan, A. Kawazoe, R.M. Goodwin, M. Conway, Y. Tateno, Q.-H. Ngo, D. Dien, A. Kawtrakul, K. Takeuchi, et al. (2008), Biocaster: detecting public health rumors with a web-based text mining system, *Bioinformatics* **24** (24), pp. 2940–2941, Oxford Univ. Press.

Culotta, A. (2010), Towards detecting influenza epidemics by analyzing Twitter messages, *Proceedings of the First Workshop on Social Media Analytics*, ACM, pp. 115–122.

Dadvar, M., C. Hauff, and F.M.G. de Jong (2011), Scope of negation detection in sentiment analysis, *Proceedings of the Dutch-Belgian Information Retrieval Workshop, DIR 2011, Amsterdam, the Netherlands*, University of Amsterdam, pp. 16–20.

Esuli, A. and F. Sebastiani (2006), SentiWordNet: A publicly available lexical resource for opinion mining, *Proceedings of LREC*, Vol. 6, pp. 417–422.

Eysenbach, G. (2006), Infodemiology: tracking flu-related searches on the web for syndromic surveillance, *AMIA Annual Symposium Proceedings*, Vol. 2006, American Medical Informatics Association, p. 244.

Fellbaum, C., editor (1998), *WordNet: An electronic lexical database*, MIT Press, Cambridge, MA.

Ferguson, N.M., D.A. Cummings, S. Cauchemez, C. Fraser, S. Riley, A. Meeyai, S. Iamsirithaworn, and D.S. Burke (2005), Strategies for containing an emerging influenza pandemic in Southeast Asia, *Nature* **437** (7056), pp. 209–214, Nature Publishing Group.

Fernández, J., E. Boldrini, J.M. Gómez, and P. Martínez-Barco (2011), Evaluating EmotiBlog robustness for sentiment analysis tasks, *Natural Language Processing and Information Systems*, Springer, pp. 290–294.

Fleiss, J.L. (1971), Measuring nominal scale agreement among many raters, *Psychological Bulletin* **76** (5), pp. 378, American Psychological Association.

Ginsberg, J., M.H. Mohebbi, R.S. Patel, L. Brammer, M.S. Smolinski, and L. Brilliant (2008), Detecting influenza epidemics using search engine query data, *Nature* **457** (7232), pp. 1012–1014, Nature Publishing Group.

Hall, M., E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I.H. Witten (2009), The WEKA data mining software: an update, *ACM SIGKDD Explorations Newsletter* **11** (1), pp. 10–18, ACM.

Kang, P. and S. Cho (2006), EUS SVMs: Ensemble of under-sampled SVMs for data imbalance problems, *Neural Information Processing*, Springer, pp. 837–846.

Landis, J.R. and G.G. Koch (1977), The measurement of observer agreement for categorical data, *biometrics* pp. 159–174, JSTOR.

Linge, J.P., R. Steinberger, T.P. Weber, R. Yangarber, E. van der Goot, D.H. Al Khudhairy, and N.I. Stilianakis (2009), Internet surveillance systems for early alerting of health threats, *Euro Surveillance* **14** (AVRJUIN), pp. 200–201, Centre Européen pour la Surveillance Épidémiologique du SIDA.

Liu, B. (2010), Sentiment analysis and subjectivity, *Handbook of Natural Language Processing,* pp. 627–666, Boca Raton, FL: CRC Press, Taylor and Francis Group.

Longini, I.M., A. Nizam, S. Xu, K. Ungchusak, W. Hanshaoworakul, D.A. Cummings, and M.E. Halloran (2005), Containing pandemic influenza at the source, *Science* **309** (5737), pp. 1083–1087, American Association for the Advancement of Science.

Mullen, T. and N. Collier (2004), Sentiment analysis using support vector machines with diverse information sources, *Proceedings of EMNLP*, Vol. 4, pp. 412–418.

Pang, B. and L. Lee (2004), A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts, *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, Association for Computational Linguistics, pp. 271–278.

Pang, B. and L. Lee (2008), Opinion mining and sentiment analysis, *Foundations and Trends in Information Retrieval* **2** (1-2), pp. 1–135, Now Publishers Inc.

Porter, M.F. et al. (1980), An algorithm for suffix stripping, *Program* **14** (3), pp. 130–137.

Prabowo, R. and M. Thelwall (2009), Sentiment analysis: A combined approach, *Journal of Informetrics* **3** (2), pp. 143–157, Elsevier.

Qiu, G., B. Liu, J. Bu, and C. Chen (2009), Expanding domain sentiment lexicon through double propagation, *Proceedings of the 21st International Joint Conference on Artificial Intelligence*, Morgan Kaufmann Publishers Inc., pp. 1199–1204.

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

Stone, P.J., D.C. Dunphy, and M.S. Smith (1966), *The General Inquirer: A Computer Approach to Content Analysis*, MIT Press.

Strapparava, C. and A. Valitutti (2004), WordNet-Affect: an affective extension of WordNet, *Proceedings of LREC*, Vol. 4, pp. 1083–1086.

Taboada, M., J. Brooke, M. Tofiloski, K. Voll, and M. Stede (2011), Lexicon-based methods for sentiment analysis, *Computational Linguistics* **37** (2), pp. 267–307, MIT Press.

Tan, S., X. Cheng, Y. Wang, and H. Xu (2009), Adapting naive Bayes to domain adaptation for sentiment analysis, *Advances in Information Retrieval* pp. 337–349, Springer.

Thacker, S.B. and R.L. Berkelman (1988), Public health surveillance in the United States, *Epidemiologic Reviews* **10** (1), pp. 164–190, Soc Epidemiolc Res.

Toutanova, K., D. Klein, C.D. Manning, and Y. Singer (2003), Feature-rich part-of-speech tagging with a cyclic dependency network, *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, Association for Computational Linguistics, pp. 173–180.

Turney, P.D. (2002), Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews, *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, Association for Computational Linguistics, pp. 417–424.

Wiebe, J. and E. Riloff (2005), Creating subjective and objective sentence classifiers from unannotated texts, *Computational Linguistics and Intelligent Text Processing* pp. 486–497, Springer.

Wilson, T., P. Hoffmann, S. Somasundaran, J. Kessler, J. Wiebe, Y. Choi, C. Cardie, E. Riloff, and S. Patwardhan (2005), OpinionFinder: A system for subjectivity analysis, *Proceedings of HLT/EMNLP on Interactive Demonstrations*, Association for Computational Linguistics, pp. 34–35.

Yang, Y. and X. Liu (1999), A re-examination of text categorization methods, *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM, pp. 42–49.

Zhang, M. and X. Ye (2008), A generation model to unify topic relevance and lexicon-based sentiment for opinion retrieval, *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM, pp. 411–418.