TASS 2015, septiembre 2015, pp 93-98 recibido 20-07-15 revisado 24-07-15 aceptado 29-07-15

Evaluating a Sentiment Analysis Approach from a Business Point of View[*]

Evaluando una aproximación de análisis de sentimientos desde un punto de vista empresarial

Javi Fernández, Yoan Gutiérrez, David Tomás,
José M. Gómez, Patricio Martínez-Barco
Department of Software and Computing Systems, University of Alicante
{javifm,ygutierrez,dtomas,jmgomez,patricio}@dlsi.ua.es

Resumen: En este artículo describimos nuestra participación a la *Tarea 1: Análisis de sentimientos a nivel global* de la competición *TASS 2015.* Este trabajo presenta la aproximación utilizada y los resultados obtenidos, enfocando la evaluación y la discusión en el contexto de las empresas de negocio.

Palabras clave: análisis de sentimientos, minería de opiniones, aprendizaje automático, Twitter

Abstract: In this paper, we describe our contribution for the *Task 1: Sentiment Analysis at global level* of the *TASS 2015* competition. This work presents our approach and the results obtained, focusing the evaluation and the discussion in the context of business enterprises.

Keywords: sentiment analysis, opinion mining, machine learning, Twitter

1 Introduction

In recent years, with the explosion of Web 2.0, textual information has become one of the most important sources of knowledge to extract useful data from. Texts can provide factual information, but also opinion-based information, such as reviews, emotions, and feelings. Blogs, forums and social networks, as well as *second screen* scenarios, offer a place for people to share information in real time. *Second screen* refers to the use of devices (commonly mobile devices) to provide interactive features on streaming content (such as television programs) provided within a software application or real-time video on social networking applications. These facts have motivated recent researches in the identification and extraction of opinions and sentiments in user comments (UC), providing invaluable information, especially for companies willing to understand customers' perceptions about their products or services in order to take appropriate business decisions. In addition, users can find opinions about a product they are interested in, and companies and personalities can monitor their online reputation.

However, processing this kind of information brings different technological challenges. The large amount of available data, its unstructured nature, and the need to avoid the loss of relevant information, makes almost impossible its manual processing. Nevertheless, *Natural Language Processing* (NLP) technologies can help in analysing these large amounts of UC automatically. Nowadays, *Sentiment Analysis* (SA) as part of an NLP task has become a popular discipline due to its wide-relatedness to social media behaviour studies. SA is commonly used to analyse the comments that people post on social networks. Also, it allows to identify the preferences and criteria of users about situations, events, products, brands, etc.

In this work we apply SA to the social context, specifically to address the *Task 1: Sentiment Analysis at global level* as part of

[*] 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 *ATTOS* (TIN2012-38536-C03-03) and *LEGOLANG* (TIN2012-31224), the European Commission by the project grant *SAM* (FP7-611312), and the University of Alicante by the project "*Explotación y tratamiento de la información disponible en Internet para la anotación y generación de textos adaptados al usuario*" (GRE13-15)

Publicado en http://ceur-ws.org/Vol-1397/. CEUR-WS.org es una publicación en serie con ISSN reconocido ISSN 1613-0073

TASS[1] 2015 challenge. This task consists on determining the global polarity of each message over provided test sets of general purpose. A detailed description about the workshop and the mentioned task can be found in (Villena-Román et al., 2015). The context of the workshop is also part of second screen phenomenon, in which users generate feedbacks of their experiences by posting them in social media. Our approach goes on that direction being part of the *SAM*[2] (Socialising Around Media) platform, where *"[...] users are interacting with media: from passive and one-way to proactive and interactive. Users now comment on or recommend a TV programme and search for related information with both friends and the wider social community."*

In this paper we present our SA system. This approach builds its own sentiment resource based on annotated samples, and based on the information collected it generates a machine learning classifier to deal with the SA challenges. The paper is structured as follows: The next section provides related works where main insights of each approach are exposed. The classification system is described in Section 3. Subsequently, Section 4 exposes in detail the evaluation, not just focusing on the guidelines of the TASS competition, but also on those aspects of interest for companies. Finally, the conclusions and future work are presented in Section 5.

2 Related Work

Different techniques have been used for both product reviews and social content analysis to obtain lexicons of subjective words with their associated polarity. We can start mentioning the strategy defined by Hu y Liu (2004) which starts with a set of seed adjectives ("good" and "bad") and reinforces the semantic knowledge by applying and expanding the lexicon with synonymy and antonymy relations provided by *WordNet*[3] (Miller, 1993). As a result, an opinion lexicon composed by a list of positive and negative opinion words for English (around 6,800 words) was obtained. A similar approach has been used for building *WordNet-Affect* (Strapparava y Valitutti, 2004) in which six basic categories of emotions (*joy*, *sadness*, *fear*, *surprise*, *anger* and *disgust*) were expanded using *WordNet*. Other widely used resource in SA is *SentiWordNet* (Esuli y Sebastiani, 2006). It was built using a set of seed words which polarity was previously known, and expanded using similarities between glosses. The main assumption behind this approach was that *"terms with similar glosses in WordNet tend to have similar polarity"*. The main problem of using these kinds of resources is that they do not consider the context in which the words appear. Some methods tried to overcome this issue building sentiment lexicons using the local context of words.

Balahur y Montoyo (2008b) built a recommender system which computed the polarity of new words using "polarity anchors" (words whose polarity is known beforehand) and *Normalised Google Distance* scores. The authors used as training examples opinion words extracted from "pros and cons reviews" from the same domain, using the clue that opinion words appearing in the "pros" section are positive and those appearing in the "cons" section are negative. Research carried out by these authors employed the lexical resource *Emotion Triggers* (Balahur y Montoyo, 2008a). Another interesting work presented by (Popescu y Etzioni, 2007) extracts the polarity from local context to compute word polarity. To this extent, it uses a weighting function of the words around the context to be classified.

In our approach, the context of the words is kept using *skipgrams*. Skipgrams are a technique whereby n-grams are formed, but in addition to allowing adjacent sequences of words, some tokens can be "skipped". The next section describes our approach in detail.

3 Methodology

Our approach is based on the one described in (Fernández et al., 2013). In this approach, the knowledge is extracted from a training dataset, where each document/sentence/tweet is labelled with respect to their overall polarity. A sentiment lexicon is created using the words, word n-grams and word skipgrams (Guthrie et al., 2006) extracted from the dataset (Section 3.1). In this lexicon, terms are statistically scored according to their appearance within each polarity (Section 3.2). Finally, a machine learning model is generated using the mentioned

[1] www.daedalus.es/TASS2015

[2] www.socialisingaroundmedia.com

[3] wordnet.princeton.edu

sentiment resource (Section 3.3). In the following sections this process is explained in detail.

3.1 Term Extraction

Each text in the dataset is processed by removing accents and converting it to lower case. Then, each text is tokenised into words, Twitter mentions (starting with `@`) and Twitter hashtags (starting with `#`). We also include combinations of punctuation symbols as terms, in order to discover some polarity-specific emoticons.

To improve the recall of our system, we perform a basic normalisation of the words extracted by removing all character repetitions. In addition, we use the stems of the words extracted, using the Snowball[4] stemmer implementation.

Afterwards, we obtain all the possible word skipgrams from those terms by making combinations of adjacent terms and skipping some of them. Specifically, we extract *k-skip-n-grams*, where the maximum number of terms in the skipgram is defined by the variable n and the maximum number of terms skipped is determined by the variable k. Note that words and word n-grams are subsets of the skipgrams extracted. Figure 1 shows an example of this process.

We must clarify the difference between two concepts: *skipgram* and *skipgram occurrence*. For example, the sentences *"I hit the tennis ball"* and *"I hit the ball"* contain the skipgram *"hit the ball"*, but there are two *occurrences* of that *skipgram*: the first one in the first example with 1 skipped term, and the second one in the second example with no skipped terms. In other words, we will consider a *skipgram* as a group of terms that appear near of each other in the same order, allowing some other terms between them, and a *skipgram occurrence* as the actual appearance of that skipgram in a text.

3.2 Term Scoring

In this step, we calculate a *global score* for each skipgram. This score using the formula in Equation 1, where T represents the set of texts in the dataset, t is a text from the dataset T, $o_{s,t}$ represents an occurrence of skipgram s in text t, and k is a function that returns the number of skipped terms of the input skipgram occurrence.

Graciaaaas por tu apoyo @usuario!! :)))
↓
Tokenisation
`Graciaaaas`, `por`, `tu`, `apoyo`, `@usuario`, `!!`, `:)))`
↓
Normalisation
`gracias`, `por`, `tu`, `apoyo`, `@usuario`, `!`, `:)`
↓
Stemming
`graci`, `por`, `tu`, `apoy`, `@usuario`, `!`, `:)`
↓
Skipgrams (2-skip-2-grams)
`graci por`, `graci tu`, `graci apoy`, `por tu`, `por apoy`, `por @usuario`, `tu apoy`, `tu @usuario`, `tu !`, `apoy @usuario`, `apoy !`, `apoy :)`, `@usuario !`, `@usuario :)`, `!  :)`

Figure 1: Term extraction process example

$$score(s) \quad = \quad \sum_{t \in T} \sum_{o_{s,t} \in t} \frac{1}{k(o_{s,t}) + 1} \qquad (1)$$

We also calculate a *polarity score* for each skipgram and polarity. It is similar to the previous score, but it only takes into account the texts with a specific polarity. The formula is presented in Equation 2, very similar to Equation 1, but where p represents a specific polarity, and T_p is the set of texts in the training corpus annotated with polarity p.

$$score(s,p) \quad = \quad \sum_{t \in T_p} \sum_{o_{s,t} \in t} \frac{1}{k(o_{s,t}) + 1} \qquad (2)$$

At the end of this process we have a list of skipgrams with a *global score* and a *polarity score*, that forms our sentiment resource.

3.3 Learning

Once we have created our statistical sentiment resource, we generate a machine learning model. We consider each polarity as a category and each text as a training instance to build our model. For each text, we will define one feature per polarity. For example, if we are categorising into *positive*, *negative* or *neutral* (3 categories), there will be 3 features for each document, called `positive`, `negative`, and `neutral` respectively.

The values for these features will be calculated using the sentiment resource, combining the previously calculated scores of all the

[4] github.com/snowballstem

$$value(p,t) = \sum_{o_{s,t} \in t} \left(\frac{1}{k(o_{s,t})+1} \cdot \frac{score(s,p)}{score(s,p)+1} \cdot \frac{score(s,p)}{score(s)}\right) \qquad (3)$$

skipgram occurrences in the text, to finally have one value for each feature. The formula used can be seen in Equation 3, where p represents a specific polarity, t is a text from the dataset, $o_{s,t}$ represents an occurrence of skipgram s in text t, and k is a function that returns the number of skipped terms of the input skipgram occurrence. This formula gives more importance to occurrences with a low number of skipped terms, with a high number occurrences in the dataset in general, and with a high number of occurrences within a specific polarity.

Finally, a model will be generated using the features specified and their values obtained as explained above. The machine learning algorithm selected is *Support Vector Machines* (SVM), due to its good performance in text categorisation tasks (Sebastiani, 2002) and previous works (Fernández et al., 2013).

4 Evaluation

Table 1 shows the official results obtained in the TASS 2015 competition, where **5L** (5 levels full test corpus), **5L1K** (5 levels 1k corpus), **3L** (3 levels full test corpus), **3L1K** (3 levels 1k corpus) represent the different datasets. **A** (accuracy), **P** (precision), **R** (recall), **F1** (F-score) represent the different measures. Finally, **Ps** (position) represents the ranking achieved in the competition. The best performance was obtained when evaluating against the 3L corpus, and the worst with the 5L1K dataset.

	A	P	R	F1	Ps
5L	0.595	0.517	0.432	0.471	12
5L1K	0.385	0.378	0.346	0.362	29
3L	0.655	0.574	0.513	0.542	14
3L1K	0.637	0.503	0.485	0.494	10

Table 1: TASS 2015 Official results

The categories specified in the workshop, `NONE` (no opinion), `P` (positive), `P+` (very positive), `N` (negative), `N+` (very negative), and `NEU` (neutral opinion) can be too granular in some cases, and specially in the context of business enterprises. Thus, we also made additional experiments using different category configurations. These are the configurations chosen:

- **Default**. In this configuration, we used the categories specified in the workshop: `NONE`, `NEU`, `P+`, `P`, `N+` and `N`.
- **Subjectivity**. In this configuration, we used only two categories: `SUBJECTIVE` and `OBJECTIVE`. The `SUBJECTIVE` includes the texts that express opinions (positive, neutral and negative), and the `OBJECTIVE` category represents no opinionated texts. The goal of this configuration is to discover users' messages that involve opinions.
- **Polarity**. In this experiment, we used only two categories: `POSITIVE` and `NEGATIVE`, independently of their intensity. The rest of the texts were discarded. By using this kind of categorisation it is possible to simplify an analysis report into only two main points of view.
- **Polarity+Neutral**. In these experiments, only the opinionated categories were used: `POSITIVE`, `NEUTRAL` and `NEGATIVE`. In this case, the `NEUTRAL` category includes both not opinionated texts and neutral text. Business companies in some cases need to consider neutral feedbacks, since the neutral mentions can also be considered as positive for their reputation.

For the experiments, we also employed additional datasets, so we can extrapolate our conclusions to other domains. Their distribution can be seen in Table 2. These are the datasets chosen:

- **TASS-Train** and **TASS-Test**. These are the official train and test dataset of the *TASS 2015 Workshop* respectively.
- **Sanders**. This is the *Sanders Dataset*[5]. It consists of hand-classified tweets labelled as *positive*, *negative* or *neutral*.

[5] www.sananalytics.com/lab/twitter-sentiment

- **MR-P**. This is the well-known *Movie Reviews Polarity Dataset 2.0*[6] (Pang y Lee, 2004). It contains reviews of movies labelled with respect to their overall sentiment polarity (*positive* and *negative*).
- **MR-PS**. The *Movie Reviews Sentence Polarity Dataset 1.0* (Pang y Lee, 2005). It has sentences from movie reviews labelled with respect their polarity (*positive* and *negative*).
- **MR-SS**. The *Movie Reviews Subjectivity Dataset 1.0* (Pang y Lee, 2004). It has sentences from movie reviews labelled with respect to their subjectivity status (*subjective* or *objective*).

These experiments were performed combining the datasets and the configurations, using *10-fold cross validation*, as these corpora do not have a default division into train and test datasets. Note that not all the datasets can be used in all configurations. For example, the *Sanders* dataset can be used to evaluate *Polarity* and *Polarity+Neutral*, but not with *Subjectivity*, as texts are not explicitly divided into not opinionated (`NONE`) and neutral (`NEU`). Table 3 shows the results obtained.

First of all, it should be noted that our model does not use information out of the training dataset. Thus, it will work very well with datasets in a specific domain and similar topics. However, in small and heterogeneous datasets the results will be lower. We consider *MR-SS*, *MR-P* and *MR-PS* as homogeneous datasets (only within the movies domain) and *TASS-Train*, *TASS-Test* and *Sanders* as heterogeneous datasets.

As we can see in Table 3, the best results were obtained in subjectivity detection in closed domains (MR-SS), with a F-score of 0.92. In open domains the results are noticeably worse. In our opinion, the results obtained are good enough for business, as studies like Wilson et al. (2005) report a 0.82 of human agreement when working with the *Polarity+Neutral* configuration.

In addition, when evaluating subjectivity the results are significantly better when the corpus is in closed domains (movies in this case), and worse in open domains. However, polarity evaluation does not seem to be as domain dependent as subjectivity evaluation. Results evaluating polarity are very similar independently of the type of dataset employed.

5 Conclusions

In this paper, we presented our contribution for the *Task 1* (Sentiment Analysis at global level) of the *TASS 2015* competition. The approach presented is a hybrid approach, which builds its own sentiment resource based on annotated samples, and generates a machine learning model based on the information collected.

Different category configurations and different data sets were evaluated to assess the performance of our approach considering business enterprises interests regarding the analysis of user feedbacks. The results obtained are promising and encourage us to continue with our research line.

As future work we plan to train our system with different datasets, in terms of size and domain, and combine our sentiment lexicon with existing ones (such as *SentiWordNet* or *WordNet Affect*) to improve the recall of our approach.

Bibliografía

Balahur, Alexandra y Andrés Montoyo. 2008a. Applying a culture dependent emotion triggers database for text valence and emotion classification. *Procesamiento del lenguaje natural*, 40:107–114.

Balahur, Alexandra y Andrés Montoyo. 2008b. Building a Recommender System using Community Level Social Filtering. En *NLPCS*, páginas 32–41.

Esuli, Andrea y Fabrizio Sebastiani. 2006. Sentiwordnet: A publicly available lexical resource for opinion mining. En *Proceedings of LREC*, volumen 6, páginas 417–422.

Fernández, Javi, Yoan Gutiérrez, José M. Gómez, Patricio Martínez-Barco, Andrés Montoyo, y Rafael Muñoz. 2013. Sentiment Analysis of Spanish Tweets Using a Ranking Algorithm and Skipgrams. En *XXIX Congreso de la Sociedad Española de Procesamiento de Lenguaje Natural (SEPLN 2013)*, páginas 133–142.

Guthrie, David, Ben Allison, Wei Liu, Louise Guthrie, y Yorick Wilks. 2006. A closer

[6] www.cs.cornell.edu/people/pabo/movie-review-data

Dataset	**NONE**	**NEU**	**P+**	**P**	**N**	**N+**	**Total**
TASS-Train	1,483	670	1,652	1,232	1,335	847	7,219
TASS-Test	21,416	1,305	20,745	1,488	11,287	4,557	60,798
Sanders	2,223		455		426		3104
MR-P	-	-	1,000		1,000		2,000
MR-PS	-	5,331		-	5,331		10,662
MR-SS	5,000	5,000					10,000

Table 2: Datasets distribution

Configuration	**Dataset**	**Accuracy**	**Precision**	**Recall**	**F1 Score**
Default	*TASS-Train*	0.810	0.446	0.337	0.383
	TASS-Test	**0.879**	**0.630**	**0.432**	**0.512**
Subjectivity	*TASS-Train*	0.806	0.770	0.628	0.692
	TASS-Test	0.740	0.717	0.693	0.705
	MR-SS	**0.925**	**0.925**	**0.925**	**0.925**
Polarity+Neutral	*TASS-Train*	0.793	0.568	0.512	0.539
	TASS-Test	**0.889**	0.708	**0.576**	**0.635**
	Sanders	0.849	**0.815**	0.492	0.614
Polarity	*TASS-Train*	0.783	0.780	0.776	0.778
	TASS-Test	**0.863**	**0.860**	**0.856**	**0.858**
	MR-P	0.825	0.825	0.825	0.825
	MR-PS	0.784	0.784	0.784	0.784
	Sanders	0.809	0.811	0.807	0.809

Table 3: Results of the evaluation of the different datasets and the different configurations

look at skip-gram modelling. En *Proceedings of the LREC-2006*, páginas 1–4.

Hu, Minqing y Bing Liu. 2004. Mining and summarizing customer reviews. En *Proceedings of the 10th ACM SIGKDD*, páginas 168–177. ACM.

Miller, George A. 1993. Five papers on WordNet. *Technical Report CLS-Rep-43, Cognitive Science Laboratory, Princeton University.*

Pang, Bo y Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. En *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, página 271.

Pang, Bo y Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. En *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, páginas 115–124.

Popescu, Ana-Maria y Orena Etzioni. 2007. Extracting product features and opinions from reviews. En *Natural language processing and text mining*. Springer, páginas 9–28.

Sebastiani, Fabrizio. 2002. Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, 34(1):1–47.

Strapparava, Carlo y Alessandro Valitutti. 2004. WordNet Affect: an Affective Extension of WordNet. En *LREC*, volumen 4, páginas 1083–1086.

Villena-Román, Julio, Janine García-Morera, Miguel A. García-Cumbreras, Eugenio Martínez-Cámara, M. Teresa Martín-Valdivia, y L. Alfonso Ureña-López. 2015. Overview of TASS 2015.

Wilson, T., P. Hoffmann, S. Somasundaran, J. Kessler, J. Wiebe, Y. Choi, C. Cardie, E. Riloff, y S. Patwardhan. 2005. OpinionFinder: A system for subjectivity analysis. En *Proceedings of HLT/EMNLP on Interactive Demonstrations.*