CEUR Workshop Proceedings

ISSN: 1613-0073

Artículos

TASS
Taller de Análisis de Sentimientos en la SEPLN

CEUR Workshop Proceedings

ISSN: 1613-0073

Organización

Comité organizador

Julio Villena-Román	Sngular	julio.villena@sngular.team
Miguel Á. García Cumbreras	Universidad de Jaén	magc@ujaen.es
Eugenio Martínez Cámara	TU Darmstadt	camara@ukp.informatik.tu-darmstadt.de
Manuel C. Díaz Galiano	Universidad de Jaén	mcdiaz@ujaen.es
M. Teresa Martín Valdivia	Universidad de Jaén	maite@ujaen.es
L. Alfonso Ureña López	Universidad de Jaén	laurena@ujaen.es

ISSN: 1613-0073
Editado en: Universidad de Jaén
Año: 2016
Editores: Julio Villena-Román Sngular julio.villena@sngular.team
Miguel Á. García Cumbreras Universidad de Jaén magc@ujaen.es
Eugenio Martínez Cámara TU Darmstadt camara@ukp.informatik.tu-darmstadt.de
Manuel C. Díaz Galiano Universidad de Jaén mcdiaz@ujaen.es
M. Teresa Martín Valdivia Universidad de Jaén maite@ujaen.es
L. Alfonso Ureña López Universidad de Jaén laurena@ujaen.es
Publicado por: CEUR Workshop Proceedings

Comité de programa

Alexandra Balahur	EC-Joint Research Centre (Italia)
José Carlos Cortizo	Universidad Europea de Madrid (España)
Jose María Gómez Hidalgo	Optenet (España)
José Carlos González-Cristobal	Universidad Politécnica de Madrid (España)
Lluís F. Hurtado	Universidad de Valencia (España)
Carlos A. Iglesias Fernández	Universidad Politécnica de Madrid (España)
Zornitsa Kozareva	Information Sciences Institute (EE.UU.)
Sara Lana Serrano	Universidad Politécnica de Madrid (España)
Ruslan Mitkov	University of Wolverhampton (Reino Unido)
Andrés Montoyo	Universidad de Alicante (España)
Rafael Muñoz	Universidad de Alicante (España)
Constantine Orasan	University of Wolverhampton (Reino Unido)
Jose Manuel Perea Ortega	Universidad de Extremadura (España)
Ferran Pla Santamaría	Universidad de Valencia (España)
María Teresa Taboada Gómez	Simon Fraser University (Canadá)
Mike Thelwall	University of Wolverhampton (Reino Unido)
José Antonio Troyano Jiménez	Universidad de Sevilla (España)

Agradecimientos

La organización de TASS ha contado con la colaboración de investigadores que participan en los siguiente proyectos de investigación:

- REDES (TIN2015-65136-C2-1-R)

TASS
Taller de Análisis de Sentimientos en la SEPLN

CEUR Workshop Proceedings

ISSN: 1613-0073

Preámbulo

Actualmente el español es la segunda lengua materna del mundo por número de hablantes tras el chino mandarín, y la segunda lengua mundial en cómputo global de hablantes. Esa segunda posición se traduce en un 6,7% de población mundial que se puede considerar hispanohablante. La presencia del español en el mundo no tiene una correspondencia directa con el nivel de investigación en el ámbito del Procesamiento del Lenguaje Natural, y más concretamente en la tarea que nos atañe, el Análisis de Opiniones. Por consiguiente, el Taller de Análisis de Sentimientos en la SEPLN (TASS) tiene como objetivo la promoción de la investigación del tratamiento del español en sistemas de Análisis de Opiniones, mediante la evaluación competitiva de sistemas de procesamiento de opiniones.

En la edición de 2016 han participado 7 equipos, de los que 6 han enviado un artículo describiendo el sistema que han presentado, habiendo sido aceptados los 6 artículos tras ser revisados por el comité organizador. La revisión se llevó a cabo con la intención de publicar sólo aquellos que tuvieran un mínimo de calidad científica.

La edición de 2016 tendrá lugar en el seno del XXXII Congreso Internacional de la Sociedad Española para el Procesamiento del Lenguaje Natural, que se celebrará el próximo mes de septiembre en Salamanca (España) dentro del V Congreso Español de Informática (CEDI 2016).

Septiembre de 2016
Los editores

TASS
Taller de Análisis de Sentimientos en la SEPLN

CEUR Workshop Proceedings

ISSN: 1613-0073

Preamble

Currently Spanish is the second native language in the world by number of speakers after the Mandarin Chinese. This second position means that the 6.7% of the world population is Spanish-speaking. The presence of the Spanish language in the world has not a direct correspondence with the number of research works related to the treatment of Spanish language in the context of Natural Language Processing, and specially in the field of Sentiment Analysis. Therefore, the Workshop on Sentiment Analysis at SEPLN (TASS) aims to promote the research of the treatment of texts written in Spanish in Sentiment Analysis systems by means of the competitive assessment of opinion processing systems.

Seven teams have participated in the 2016 edition of the workshop. Six of the seven teams have submitted a description paper of their systems. After a review process, the organizing committee has accepted the 6 papers, because all of them reached an acceptable scientific quality level.

The 2016 edition will be held at the 32nd International Conference of the Spanish Society for Natural Language Processing (SEPLN 2016), which will take place at Salamanca in September framed by the 5th Spanish Conference of Computer Science (CEDI 2016).

September 2016
The editors

 ISSN 1613-0073

CEUR Workshop Proceedings

ISSN: 1613-0073

Artículos

Artículos

Overview of TASS 2016

Resumen de TASS 2016

Miguel Ángel García Cumbreras[1], Julio Villena Román[2], Eugenio Martínez Cámara[1], Manuel Carlos Díaz Galiano[1], M. Teresa Martín Valdivia[1], L. Alfonso Ureña López[1]
[1]Universidad de Jaén
23071 Jaén, Spain
[2]Sngular
28034 Madrid, Spain
[1]{magc, emcamara, mcdiaz, laurena, maite}@ujaen.es [2]{julio.villena}@sngular.team

Resumen: Este artículo describe la quinta edición del taller de evaluación experimental TASS 2016, enmarcada dentro del Congreso Internacional SEPLN 2016. El principal objetivo de TASS es promover la investigación y el desarrollo de nuevos algoritmos, recursos y técnicas para el análisis de sentimientos en medios sociales (concretamente en Twitter), aplicado al idioma español. Este artículo describe las tareas propuestas en TASS 2016, así como el contenido de los corpus utilizados, los participantes en las distintas tareas, los resultados generales obtenidos y el análisis de estos resultados.
Palabras clave: TASS 2016, análisis de opiniones, medios sociales

Abstract: This paper describes TASS 2016, the fifth edition of the Workshop on Sentiment Analysis at SEPLN. The main aim is the promotion of the research and the development of new algorithms, resources and techniques on the field of sentiment analysis in social media (specifically Twitter) focused on the Spanish language. This paper presents the TASS 2016 proposed tasks, the description of the corpora used, the participant groups, the results and analysis of them.
Keywords: TASS 2016, sentiment analysis, social media.

1 Introduction

TASS is an experimental evaluation workshop, a satellite event of the annual SEPLN Conference, with the aim to promote the research on Sentiment Analysis in social media focused on the Spanish language. The fifth edition will be held on September 13th, 2016 at the University of Salamanca, Spain.

Sentiment Analysis (SA) is traditionally defined as the computational treatment of opinion, sentiment and subjectivity in texts (Pang & Lee, 2008). However, Cambria and Hussain (2012) offer a more updated definition: Computational techniques for the extraction, classification, understanding and evaluation of opinions and comments published on the Internet and other kind of user generated contents. It is a hard task because even humans often disagree on the polarity of a given text. And it is a harder task when the text has only 140 characters (Twitter messages or tweets).

Although SA is not a new task, it is still challenging, because the state of the art has not yet resolved some problems related to multilingualism, domain adaptation, text genre adaptation and polarity classification at fine grained level. Polarity classification has usually been tackled following two main approaches. The first one applies machine learning algorithms in order to train a polarity classifier using a labelled corpus (Pang et al. 2002). This approach is also known as the supervised approach. The second one is known as semantic orientation, or the unsupervised approach, and it integrates linguistic resources in a model in order to identify the valence of the opinions (Turney 2002).

The aim of TASS is to provide a competitive forum where the newest research works in the field of SA in social media, specifically focused on Spanish tweets, are described and discussed by scientific and business communities.

The rest of the paper is organized as follows. Section 2 describes the different corpus

provided to participants. Section 3 shows the different tasks of TASS 2016. Section 4 describes the participants and the overall results are presented in Section 5. Finally, the last section shows some conclusions and future directions.

2 Corpus

TASS 2016 experiments are based on two corpora, specifically built for the different editions of the workshop.

The two corpora will be made freely available to the community after the workshop. Please send an email to tass@sngularmeaning.team filling in the TASS Corpus License agreement with your email, affiliation (institution, company or any kind of organization) and a brief description of your research objectives, and you will be given a password to download the files in the password protected area. The only requirement is to include a citation to a relevant paper and/or the TASS website.

2.1 General corpus

The General Corpus contains over 68.000 tweets, written in Spanish, about 150 well-known personalities and celebrities of the world of politics, economy, communication, mass media and culture, between November 2011 and March 2012. Although the context of extraction has a Spanish-focused bias, the diverse nationality of the authors, including people from Spain, Mexico, Colombia, Puerto Rico, USA and many other countries, makes the corpus reach a global coverage in the Spanish-speaking world.

Each tweet includes its ID (*tweetid*), the creation date (*date*) and the user ID (*user*). Due to restrictions in the Twitter API Terms of Service (https://dev.twitter.com/terms/api-terms), it is forbidden to redistribute a corpus that includes text contents or information about users. However, it is valid if those fields are removed and instead IDs (including Tweet IDs and user IDs) are provided. The actual message content can be easily obtained by making queries to the Twitter API using the *tweetid*.

The general corpus has been divided into training set (about 10%) and test set (90%). The training set was released, so the participants could train and validate their models. The test corpus was provided without any tagging and has been used to evaluate the results. Obviously, it was not allowed to use the test data from previous years to train the systems.

Each tweet was tagged with its global polarity (positive, negative or neutral sentiment) or no sentiment at all. A set of 6 labels has been defined: strong positive (P+), positive (P), neutral (NEU), negative (N), strong negative (N+) and one additional no sentiment tag (NONE).

In addition, there is also an indication of the level of agreement or disagreement of the expressed sentiment within the content, with two possible values: AGREEMENT and DISAGREEMENT. This is especially useful to make out whether a neutral sentiment comes from neutral keywords or else the text contains positive and negative sentiments at the same time.

Moreover, the polarity values related to the entities that are mentioned in the text are also included for those cases when applicable. These values are similarly tagged with 6 possible values and include the level of agreement as related to each entity.

This corpus is based on a selection of a set of topics. Thematic areas such as “política” (“politics”), “fútbol” (“soccer”), “literatura” (“literature”) or “entretenimiento” (“entertainment”). Each tweet in the training and test set has been assigned to one or several of these topics (most messages are associated to just one topic, due to the short length of the text).

The annotation has been semi-automatically done: a baseline machine learning model is first run and then all tags are checked by human experts. In the case of the polarity at entity level, due to the high volume of data to check, the human annotation has only been done for the training set.

Table 1 shows a summary of the training and test corpora provided to participants.

Attribute	Value
Tweets	68.017
Tweets (test)	60.798 (89%)
Tweets (test)	7.219 (11%)
Topics	10
Users	154
Date start (train)	2011-12-02
Date end (train)	2012-04-10
Date start (test)	2011-12-02
Date end (test)	2012-04-10

Table 1: Corpus statistics

Users were journalists (*periodistas*), politicians (*políticos)* or celebrities (*famosos*). The only language involved was Spanish (*es)*.

The list of topics that have been selected is the following:

- Politics (política)
- Entertainment (entretenimiento)
- Economy (economía)
- Music (música)
- Soccer (fútbol)
- Films (películas)
- Technology (tecnología)
- Sports (deportes)
- Literature (literatura)
- Other (otros)

The corpus is encoded in XML. Figure 1 shows the information of two tweets. The first tweet is only annotated with the polarity at tweet level because there is not any entity in the text. However, the second one is annotated with the global polarity of the message and the polarity associated to each of the entities that appear in the text (UPyD and Foro Asturias).

```
<tweet>
    <tweetid>0000000000</tweetid>
    <user>usuario0</user>
    <content>
        <![CDATA['Conozco a alguien q es adicto al drama! Ja ja ja te suena d algo!]]>
    </content>
    <date>2011-12-02T02:59:03</date>
    <lang>es</lang>
    <sentiments>
        <polarity>
            <value>P+</value>
            <type>AGREEMENT</type>
        </polarity>
    </sentiments>
    <topics>
        <topic>entretenimiento</topic>
    </topics>
</tweet>
<tweet>
    <tweetid>0000000001</tweetid>
    <user>usuario1</user>
    <content>
        <![CDATA['UPyD contará casi seguro con grupo gracias al Foro Asturias.]]>
    </content>
    <date>2011-12-02T00:21:01</date>
    <lang>es</lang>
    <sentiments>
        <polarity>
            <value>P</value>
            <type>AGREEMENT</type>
        </polarity>
        <polarity>
            <entity>UPyD</entity>
            <value>P</value>
            <type>AGREEMENT</type>
        </polarity>
        <polarity>
            <entity>Foro_Asturias</entity>
            <value>P</value>
            <type>AGREEMENT</type>
        </polarity>
    </sentiments>
  - <topics>
        <topic>política</topic>
    </topics>
</tweet>
```

Figure 1: Sample tweets (General corpus)

2.2 STOMPOL corpus

STOMPOL (corpus of Spanish Tweets for Opinion Mining at aspect level about POLitics) is a corpus of Spanish tweets prepared for the research on the challenging task of opinion mining at aspect level. The tweets were gathered from 23rd to 24th of April 2015, and are related to one of the following political aspects that appear in political campaigns:

- Economics (Economía): taxes, infrastructure, markets, labour policy...
- Health System (Sanidad): hospitals, public/private health system, drugs, doctors...
- Education (Educación): state school, private school, scholarships...
- Political party (Propio_partido): anything good (speeches, electoral programme...) or bad (corruption, criticism) related to the entity
- Other aspects (Otros_aspectos): electoral system, environmental policy...

Each aspect is related to one or several entities that correspond to one of the main political parties in Spain, which are:

- Partido_Popular (PP)
- Partido_Socialista_Obrero_Español (PSOE)
- Izquierda_Unida (IU)
- Podemos
- Ciudadanos (C's)
- Unión_Progreso_y_Democracia (UPyD)

Each tweet in the corpus has been manually annotated by two annotators, and a third one in case of disagreement, with the sentiment polarity at aspect level. Sentiment polarity has been tagged from the point of view of the person who writes the tweet, using 3 levels: P, NEU and N. Again, no difference is made between no sentiment and a neutral sentiment (neither positive nor negative). Each political aspect is linked to its correspondent political party and its polarity.

Figure 2 shows the information of two sample tweets.

```
<tweet id="591267548311769088">@ahorapodemos @Pablo_Iglesias_ @SextaNocheTV
Que alguien pregunte si habrá cambios en las <sentiment aspect="Educacion"
entity="Podemos" polarity="NEU">becas</sentiment> MEC para universitarios, por
favor.</tweet>

<tweet id="591192167944736769">#Arroyomolinos lo que le interesa al ciudadano
son Políticos cercanos que se interesen y preocupen por sus problemas <
sentiment aspect="Propio_partido" entity="Union_Progreso_y_Democracia" polarity
="P">@UPyD</sentiment> VECINOS COMO TU</tweet>
```

Figure 2: Sample tweets (STOMPOL corpus)

The number of tweets per each entity are shown in Table 2.

Entity	Train	Test
PP	205	125
PSOE	136	70
C's	119	87
Podemos	98	80
IU	111	43
UPyD	97	124
Total	766	529

Table 2: Number of tweets per entity and per corpus subset

3 Description of tasks

Since the first edition of TASS, a new task and a new corpus have been published. However, one of the aims of TASS is the evaluation of the progress of the research on SA. Thus, the edition of 2016 was focused on the analysis and the comparison of the systems with the submissions of previous editions.

The edition of 2016 was focused on two tasks: polarity classification at tweet level and polarity classification at entity level. The polarity classification task has been proposed with the same corpus since the first edition of TASS, but the polarity classification at aspect level has been proposed with a different corpus each edition. In the edition of 2016 the classification at aspect level uses the STOMPOL corpus, which was published the first time in the edition of 2015.

Participants are expected to submit up to 3 results of different experiments for one or both of these tasks, in the appropriate format described below.

Along with the submission of experiments, participants have been invited to submit a paper to the workshop in order to describe their experiments and discussing the results with the audience in a regular workshop session.

The two proposed tasks are described next.

3.1 Task 1: Sentiment Analysis at Global Level

This task consists on performing an automatic polarity classification to determine the global polarity of each message in the test set of the General Corpus. The training set of the corpus was provided to the participants with the aim they could train and validate their models with it. There were two different evaluations: one based on 6 different polarity labels (`P+`, `P`, `NEU`, `N`, `N+`, `NONE`) and another based on just 4 labels (`P`, `N`, `NEU`, `NONE`).

Participants are expected to submit (up to 3) experiments for the 6-labels evaluation, and they are also allowed to submit (up to 3) specific experiments for the 4-labels scenario.

Results must be submitted in a plain text file with the following format:

```
tweetid \t polarity
```

where polarity can be:

- `P+`, `P`, `NEU`, `N`, `N+` and `NONE` for the 6-labels case
- `P`, `NEU`, `N` and `NONE` for the 4-labels case.

The same test corpus of previous years was used for the evaluation in order to develop a comparison among the systems. The accuracy is one of the measures used to evaluate the systems, however due to the fact that the training corpus is not totally balanced the systems were also assessed by the macro-averaged precision, macro-averaged recall and macro-averaged F1-measure.

3.2 Task 2: Aspect-based sentiment analysis

A corpus with the entities and the aspect identified was provided to the participants, so the goal of the systems is the inference of the polarity at the aspect-level. As in 2015, STOMPOL corpus was the corpus used in this task. STOMPOL was divided in training and test set, the first one for the development and validation of the systems, and the second for evaluation.

Participants are expected to submit up to 3 experiments for each corpus, each in a plain text file with the following format:

```
tweetid \t aspect-entity \t polarity
```

Allowed polarity values are: P, N and NEU. For the evaluation, a single label combining "aspect-polarity" has been considered. As in the first task, accuracy, macro-averaged precision, macro-averaged recall and macro-averaged F1-measure have been calculated for the global result.

4 Participants and Results

This year 7 (7 last year) groups submitted their systems The list of active participant groups is

shown in Table 3, including the tasks in which they have participated.

Six of the seven participant groups sent a report describing their experiments and results achieved. Papers were reviewed and included in the workshop proceedings. References are listed in Table 4.

Group	1	2
jacerong	X	
ELiRF-UPV	X	X
LABDA	X	
INGEOTEC	X	
GASUCR	X	
GTI		X
SINAI_w2v	X	
Total	6	1

Table 3: Participant groups

Group	Report
ELiRF	ELiRF-UPV en TASS 2016: Análisis de Sentimientos en Twitter
GTI	GTI at TASS 2016: Supervised Approach for Aspect Based Sentiment Analysis in Twitter
jacerong	JACERONG at TASS 2016: An Ensemble Classifier for Sentiment Analysis of Spanish Tweets at Global Level
LABDA	LABDA at the 2016 TASS challenge task: using word embedding for the sentiment analysis task
SINAI	Participación de SINAI en TASS 2016

Table 4: Participant reports

5 Results

This section will be focused on the description and the analysis of the results and the systems submitted by the participants.

5.1 Task 1: Sentiment Analysis at Global Level

Submitted runs and results for Task 1, evaluation based on 5 polarity levels with the whole General test Corpus are shown in Table 5. Accuracy, macro-averaged precision, macro-averaged recall and macro-averaged F1-measure have been used to evaluate each individual label and ranking the systems.

Run Id	M-F1
ELiRF-UPV_1	0.518
jacerong_2	0.504
jacerong_3	0.503
jacerong_1	0.499
ELiRF-UPV_2	0.496
INGEOTEC	0.464
LABDA_1	0.429
LABDA_2	0.429
LABDA_3	0.418
GASURC_3	0.254
GASURC_1	0.232
GASURC_2	0.227

Table 5: Results for Task 1, 5 levels

In order to perform a more in-depth evaluation, results are calculated considering the classification only in 3 levels (POS, NEU, NEG) and no sentiment (NONE) merging P and P+ in only one category, as well as N and N+ in another one. The results reached by the submitted systems are shown in Table 6.

Run Id	M-F1
jacerong_3	0.568
jacerong_2	0.567
jacerong_1	0.564
ELiRF-UPV_1	0.549
ELiRF-UPV_2	0.548
INGEOTEC	0.524
LABDA_3	0.511
LABDA_2	0.508
LABDA_1	0.508
SINAI_w2v_1	0.504
SINAI_w2v_3	0.486
SINAI_w2v_4	0.469
SINAI_w2v_2	0.440
GASURC_1	0.250
GASURC_2	0.152

Table 6: Results for Task 1, 3 levels

5.2 Task 2: Aspect-based Sentiment Analysis

Submitted runs and results for Task 2, with the STOMPOL corpus, are shown in Table 7. Accuracy, macro-averaged precision, macro-averaged recall and macro-averaged F1-measure have been used to evaluate each individual label and ranking the systems.

Run Id	M-F1
ELiRF-UPV_1	0.526
GTI	0.463

Table 7: Results for Task 2

5.3 Description of the systems

The systems submitted in the edition of 2016 represent the next step of the ones submitted in the previous edition. The systems may be cluster in two groups, those ones that rely on the classification power of the ensemble of several base classifiers, and those systems that change the use traditional Bag-of-Words model for the use of vectors of word embeddings in order to represent the meaning of each word. In the subsequent paragraphs the main features of the systems submitted are going to be depicted.

Hurtado and Pla (2016) describe the participation of the team ELiRF-UPV in the two tasks of TASS 2016. The only difference between the systems submitted for the two tasks is the fact that the one focused on the second task has a module for the identification of the context of each of the entities and aspects annotated on the tweets. The polarity classification system relies on the ensemble of 192 configurations of a SVM classifiers. For the combination of the set of classifiers they evaluate the performance of an approach based on voting and other on stacking.

The system depicted in (Cerón-Guzmán, 2016) is also based on an approach of ensemble classifiers. In this case the base classifiers used a classifier based on logistic regression and they are combined by voting.

Alvarez et al. (2016) exposed the participation of the team GTI on the task 2. The system is similar to the system of the team ELiRF-UPV in the sense that it is composed by two layers: context identification and polarity classification. Regarding the identification of the context, the authors design a heuristic method based on lexical markers. The polarity classification system is a SVM classifier that uses different type of features in order to represent the contexts of the entities and the aspects.

Montejo-Ráez and Díaz-Galiano (2016) introduce a system based on a supervised learning algorithm over vectors resulting from a weighted vector. This vector is computed using a Word2Vec algorithm. This method, which is inspired from neural-network language modelling, was executed with a collection of tweets written in Spanish and the Spanish Wikipedia in order to generate a set of word embeddings for the representation of the words of the General Corpus of TASS as dense vectors. The creation of the collection of tweets written in Spanish followed a distant supervision approach by means the assumption that tweets with happy and sad emoticons express emotions or opinions. Their experiments show massive data from Twitter can lead to a slight improvement in classification accuracy.

The system presented by the team LABDA (Quirós, Segura-Bedmar and Paloma Martínez, 2016) is similar to the one submitted by SINAI (Montejo-Ráez and Díaz-Galiano, 2016) because it also used word embeddings as schema of representation of the meaning of the words of the tweets. Quirós, Segura-Bedmar and Paloma Martínez (2016) assessed the performance of the SVM and Logistic Regression as classifiers.

Casasola Murillo and Marín Reventós (2016) submitted an unsupervised system based on the system described in Turney (2002), but with a specific adaptation to the classification of tweets written in Spanish.

5.4 Analysis

In Table 5 and Table 6 are shown the results of each system and they are ranked by the F1-score reached, so it is not hard to know what is the best system in the edition of 2016.

On the other hand, how many tweets were rightly classified by the submitted systems? Is there a set of tweets that were not rightly classified by any system? What are the most difficult tweets to classify? These questions are going to be answered in the following paragraphs?

Table 8 shows the rate of tweets that are rightly classified by a number of systems. There

are about a 6% of tweets whose polarity is not inferred by any of the submitted systems. In other words, the submitted systems in the edition of 2016 are able to classify about the 94% of the test set. So, what is the main features of that 6% of tweets that any system inferred their polarity?

Number of systems	Rate of tweets
0	0.056%
1	0.065%
2	0.063%
3	0.067%
4	0.059%
5	0.061%
6	0.074%
7	0.078%
8	0.081%
9	0.112%
10	0.122%
11	0.082%
12	0.062%
13	0.011%

Table 8: Rate of tweets rightly classified (6 classes) by a number of systems

Id: 171304000392663040

Sacarle 17 puntos en la final de Copa al Barça CB en el Palau Sant Jordi es una pasada.

Beating Barça by 17 points in the Copa is amazing

Polarity: P+

Figure 3: Tweet not rightly classified by any system

Figures Figure 3,Figure 4Figure 5 are three examples of tweets that were not rightly classified by any system. The common feature of the three tweets is that they do not have any lexical marker that express emotion or opinion. Moreover, the tweet of the Figure 4 is sarcastic, which means an additional challenging for SA because requires a deep understanding of the language.

Id: 177439342497767424

hahahahahaha "@Absolutexe: ¿Le han cambiado ya el nombre a la Junta de Andalucía por la Banda de Andalucía o aún no?"

hahahahahaha "@Absolutexe: Has the Junta de Andalucía renamed Gang of Andalucía or not yet?"

Polarity: N+

Figure 4: Tweet not rightly classified by any system

Id: 177439342497767424

Rubalcaba pide a Rajoy que presente ya los Presupuestos y dice que no lo hace porque espera a las elecciones andaluzas

Rubalcaba requires Rajoy to submit the Budget and says that he didn't because he is waiting the results of the elections in Andalucia

Polarity: NONE

Figure 5: Tweet not rightly classified by any system

All the systems submitted are based on linear classifiers that do not take into account the context of each word, which means a big drawback for the understanding the meaning of a span of text.

The tweets of the Figures 3, 4 and 5 show that opinions and emotions are not only expressed by lexical markers, so the future participants should take into account the challenging task of implicit opinion analysis, irony and sarcasm detection. These new problems may be framed on the semantic level of Natural Language Processing and should be tackled by the research community in order to go a step further in the understanding of the subjective information, which is continuously published on the Internet.

6 Conclusions and Future Work

TASS was the first workshop about SA focused on the processing of texts written in Spanish. In the three first editions of TASS, the research community were mainly formed by Spanish researchers, however since the last edition, the researchers that come from South America is making bigger, so it is an evidence that the research community of Sentiment Analysis in Spanish is not only located in Spain and is formed by the Spanish speaking countries.

Anyway, the developed corpus and gold standards, and the reports from participants will for sure be helpful for knowing the state of the art in SA in Spanish.

The future work will be mainly focused on the definition of a new General Corpus because of the following reasons:

1. The language used on Twitter changes faster than the language used in traditional genres of texts, so the update of the corpus is required in order to cover a real used of the language on Twitter.
2. After several editions of the workshop, we realize that the quality of the annotation is not extremely good, so it is required to define a new corpus with a high quality annotation in order to provide a real gold standard for Spanish SA on Twitter.
3. The research community deeply know the General Corpus of TASS and it wants a new challenge.

A significant amount of new tasks is currently being defined in Natural Language Processing, so some of them, such as stance classification, will be studied to be proposal for the next edition of TASS.

Acknowledgements

This work has been partially supported by a grant from the Fondo Europeo of Desarrollo Regional (FEDER) and REDES project (TIN2015-65136-C2-1-R) from the Spanish Government.

References

Cambria, E. and Amir Hussain, A. 2012. Sentic Computing. Techniques, Tools and Applications. Springer Briefs in Cognitive Computation, volume 2. Springer Netherlands. ISBN 978-94-007-5069-2. doi:10.1007/978-94-007-5070-8.

Cerón-Guzmán, J. A. 2016. JACERONG at TASS 2016: An Ensemble Classifier for Sentiment Analysis of Spanish Tweets at Global Level. In Proceedings of TASS 2016: Workshop on Sentiment Analysis at SEPLN co-located with the 32nd SEPLN Conference (SEPLN 2016), Salamanca, September

Casasola Murillo, E. and Gabriela M. R. 2016. Evaluación de Modelos de Representación del Texto con Vectores de Dimensión Reducida para Análisis de Sentimiento. In Proceedings of TASS 2016: Workshop on Sentiment Analysis at SEPLN co-located with the 32nd SEPLN Conference (SEPLN 2016), Salamanca, September

Hurtado, Ll. and Ferran P. 2016. ELiRF-UPV en TASS 2016: Análisis de Sentimientos en Twitter. In Proceedings of TASS 2016: Workshop on Sentiment Analysis at SEPLN co-located with the 32nd SEPLN Conference (SEPLN 2016), Salamanca, September

Montejo-Ráez, A. and Díaz-Galiano, M. C. 2016. Participación de SINAI en TASS 2016. In Proceedings of TASS 2016: Workshop on Sentiment Analysis at SEPLN co-located with the 32nd SEPLN Conference (SEPLN 2016), Salamanca, September

Pang, B., Lillian Lee and Shivakumar Vaithyanathan. 2002. Thumbs up?: Sentiment classification using machine learning techniques. In Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10, EMNLP '02, páginas 79–86. Association for Computational Linguistics, Stroudsburg, PA, USA. doi:10.3115/1118693.1118704.

Pang, B. and Lillian Lee (2008). Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, 2(1-2):1–135. ISSN 1554-0669. doi:10.1561/1500000011.

Quirós, A., Isabel S. B. and Paloma M. 2016. LABDA at the 2016 TASS challenge task: using word embeddings for the sentiment analysis task. In Proceedings of TASS 2016: Workshop on Sentiment Analysis at SEPLN co-located with the 32nd SEPLN Conference (SEPLN 2016), Salamanca, September

Turney, P. D. 2002. Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews. In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, ACL '02, pp: 417–424. Association for Computational Linguistics, Stroudsburg, PA, USA. doi:10.3115/1073083.1073153.

Villena-Román, J., Sara, L. S., Eugenio M. C., and José Carlos G. C. 2013. *TASS - Workshop on Sentiment Analysis at SEPLN*. Revista de Procesamiento del Lenguaje Natural, 50, pp 37-44.

Villena-Román, J., Janine G. M., Sara L. S. and José Carlos G. C. 2014. *TASS 2013 - A Second Step in Reputation Analysis in Spanish*. Revista de Procesamiento del Lenguaje Natural, 52, pp 37-44.

TASS 2016: Workshop on Sentiment Analysis at SEPLN, septiembre 2016, pág. 23-28

Evaluación de Modelos de Representación del Texto con Vectores de Dimensión Reducida para Análisis de Sentimiento*

Evaluation of Reduced Dimension Vector Text Representation Models for Sentiment Analysis

Edgar Casasola Murillo
Universidad de Costa Rica
San José, Costa Rica
edgar.casasola@ucr.ac.cr

Gabriela Marín Raventós
Universidad de Costa Rica
San José, Costa Rica
gabriela.marin@ucr.ac.cr

Resumen: Se describe el sistema para análisis de sentimiento desarrollado por el Grupo de Análisis de Sentimiento GAS-UCR de la Universidad de Costa Rica para la tarea 1 del workshop TASS 2016. El sistema propuesto está basado en el uso de vectores de características de baja dimensión para representación del texto. Se propone un modelo simple fundamentado en la normalización de texto con identificación de marcadores de énfasis, el uso de modelos de lenguaje para representar las características locales y globales del texto, y características como emoticones y partículas de negación. Los primeros experimentos muestran las mejoras que se obtienen en la precisión al identificar la polaridad de textos completos conforme se van incorporando las características aquí mencionadas.
Palabras clave: análisis de sentimiento, clasificación de textos por polaridad, textos cortos

Abstract: The Sentiment Analisys System developed by GAS-UCR team of the University of Costa Rica for task 1 of TASS 2016 workshop is presented. Preliminar evaluation results of the proposed Sentiment Analysis System are presented. The system is based on low dimension feature vectors for text representation. The proposed model is based on text normalization with emphasis mark identification, the use of local and global language models, and other features like emoticons an negation terms. Initial experimentation shows that the introduction of the selected features have a positive impact on precision at the polarity classification task.
Keywords: sentiment analysis, polarity based text clasification, short texts.

1 Introducción

Este trabajo tiene como propósito describir el sistema utilizado por el grupo de investigación en análisis de sentimiento de la Universidad de Costa Rica en su participación en el taller TASS2016 (García-Cumbreras et al., 2016). El enfoque del trabajo del grupo ha sido el estudio de los factores que van incidiendo en las mejoras en la precisión obtenida al llevar a cabo la clasificación de la polaridad de *tweets* en idioma español. Nuestro sistema se fundamenta en tres elementos básicos que son: la normalización del texto en la etapa de preprocesamiento identificando los potenciales marcadores de énfasis presentes en el mismo, la creación de vectores de características de dimensión reducida para disminuir el efecto de la dispersión de los datos, y la exploración del impacto del uso de diccionarios de polaridad que se generan mediante la utilización de diferentes modelos de representación del lenguaje asociados tanto al contexto local como global de los datos. Para esto estamos utilizando una adaptación propia del algoritmo de Turney (Turney, 2002)sobre un corpus de 5 millones de *tweets* en español. Estos modelos se almacenan en forma de diccionarios con polaridad para su posterior reutilización. Nos interesa particularmente la investigación en este campo dado que si bien desde el año 2013 se identificó una brecha importante entre la cantidad de investigación y tecnología del lenguaje desarrollada para el

* Este trabajo se ha llevado a cabo gracias al apoyo económico de la Universidad de Costa Rica y el Gobierno de la República de Costa Rica a través del MICITT. Se agradece a los asistentes del grupo de investigación GAS-UCR por su trabajo

idioma inglés y el español (Cambria et al., 2013) (Melero et al., 2012), de la misma forma debemos tener presente que no necesariamente las soluciones para español peninsular van a tener los mismos resultados al aplicarse a variantes de español americano, por lo que los recursos y métodos que utilizamos tienen la intensión de aportar a la investigación en español y colaborar para su posterior aplicación en otros contextos de habla hispana.

2 Antecedentes

Entre los resultados obtenidos con sistemas con enfoques basados en aprendizaje máquina, el uso de **máquina de soporte vectorial (MSV)** ha ofrecido buenos resultados tanto en inglés (Kiritchenko, Zhu, y Mohammad, 2014) y (Batista y Ribeiro, 2013) como en español donde 9 de los 14 sistemas para el español presentados en TASS2015 (Villena-Román et al., 2015) hacían uso de este tipo de clasificador. Sin embargo, la dependencia del lenguaje hace que estos clasificadores dependan de los vectores de características con los que son representados los comentarios de texto. Esta extracción de características ha sido el foco de atención de múltiples trabajos como (Cabanlit y Junshean Espinosa, 2014) , (Feldman, 2013), (Guo y Wan, 2012), (Sharma y Dey, 2012) y (Wang et al., 2011). En trabajos recientes de análisis de sentimiento en español tales como el trabajo de (Martínez-Cámara et al., 2015) se utilizan varios diccionarios de polaridad y se representan utilizando un modelo de espacio vectorial MEV. El diccionario en sí se convierte en un modelo de lenguaje que sirve como recurso para lograr representaciones eficientes de los vectores utilizados para la clasificación.

En los últimos años la representación vectorial basada en modelos de lenguaje como unigramas y bigramas se movió hacia representaciones de características ya que la cantidad de términos introduce un problema asociado a su alta dispersión en el vector (Cambria et al., 2013). Si los vectores contienen un alto número de atributos diferentes, uno por término, los conjuntos de datos para entrenamiento deben contener una mayor cantidad de textos anotados que atributos para un buen entrenamiento de los clasificadores. Es por esto que los modelos de representación del lenguaje basados en unigramas, bigramas o bien skipgramas requiren de una representación vectorial eficiente. Trabajos recientes buscan la representación vectorial de las palabras en el espacio continuo como es el caso del uso de Word2Vect (Díaz-Galiano y Montejo-Ráez, 2015).

3 Descripción del sistema

Nuestro sistema se fundamenta en cuatro elementos que consideramos importantes de mencionar. Primero nos referiremos a la forma en que construimos nuestro diccionario con la polaridad de los términos y las razones para haber construido uno propio. Posteriormente nos referimos a nuestro proceso de preprocesamiento e identificación de potenciales marcadores de énfasis durante esta etapa inicial. En la siguiente subsección explicamos la forma en que construimos vectores de baja dimensión con información y hacemos uso del diccionario. Finalmente se menciona la forma en que se pretende capturar en los vectores de características aspectos locales con respecto a los datos de entrenamiento, y globales, a partir de modelos de representación del lenguaje general.

3.1 Creación del diccionario polarizado

Decidimos desarrollar diccionarios de polaridad propios, en lugar de utilizar los existentes, ya que consideramos que desde el punto de vista del procesamiento de lenguaje natural tradicional (Indurkhya y Damerau, 2010) estos diccionarios con polaridad pueden ser vistos cada uno, como un modelo de lenguaje particular. Por este motivo tratamos de desarrollar y evaluar una adaptación del tradicional método de generación de estos recursos lingüísticos de (Turney, 2002). La decisión anterior no se debió a la no existencia de diccionarios polarizados ya que claramente en trabajos como (Martínez-Cámara et al., 2015) se hace uso de varios de ellos, sino con el fin de incorporar la etapa de creación de diccionario dentro de la metodología de trabajo para que posteriores investigaciones en otros países de habla hispana puedan replicar el trabajo y disminuir la barrera inicial asociada a la falta de recursos lingüísticos propios y el efecto del uso del diccionario polarizado sobre la calidad de los resultados de clasificación.

El diccionario de polaridad creado utiliza un corpus recolectado durante el año 2013, con 5 millones de *tweets* en español. La variante con respecto al algoritmo propuesto

por Turney (Turney, 2002) es la siguiente. Para el cálculo de la **orientación semántica de un término**, tal y como lo define Turney en su artículo original, se utilizaron grupos de palabras semilla en lugar de un solo término, y en lugar de utilizar consultas a motores de búsqueda para obtener la cantidad de textos donde aparecen las palabras analizadas cerca de las palabras positivas o negativas se utilizó el motor de búsqueda implementado con el software libre Solr `http://lucene.apache.org/solr/`. Con el motor se indexaron los 5 millones de *tweets* por lo que las consultas se ejecutaron en forma local. Este método cuenta con la ventaja de que se puede calcular entonces la orientación semántica de un término directamente o bien almacenarlo en un diccionario. En nuestro caso precalculamos la polaridad y la almacenamos en forma de diccionario. Por el momento solo se han llevado a cabo los cálculos para términos individuales.

3.2 Normalizador de texto con marcadores de énfasis

Luego de un proceso de análisis de las características presentes en el texto desarrollamos un sistema para normalización del texto. Para este preprocesamiento se segmentan los términos potenciales, signos de puntuación y emoticones. Se lleva a cabo un marcado y conversión de los términos. El proceso que seguimos hace una eliminación de los términos que son identificados en el diccionario. Este proceso se muestra en la figura 1.

Las repeticiones de letras, repeticiones de sílabas y mayúsculas son identificadas y eliminadas pero estos términos se marcan como potenciales identificadores de énfasis. Ejemplos son: **EXCELENTE**, **graciassss**, **buenisísimo**. En esta fase se identifican los *tweets* que contienen palabras positivas con énfasis para su posterior uso.

3.3 Representación vectorial de baja dimensión

Dos características representadas en los vectores tienen que ver con la presencia y polaridad de los emoticones y con la presencia de partículas de negación. Además, al desarrollar esta investigación se pudo observar que los términos positivos con marcadores de énfasis son un potencial identificador de la polaridad positiva de los textos que los contienen, por lo tanto esta característica también fue incorporada. La presencia de marcadores de énfasis tales como repetición de caracteres, de sílabas, o mayúsculas sobre términos que aparecen como negativos en algún contexto son registrados como una característica importante en el vector.

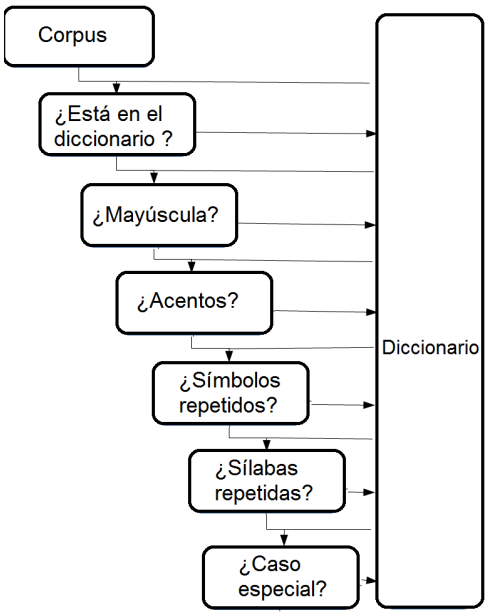

Figura 1: Proceso de normalizacion del texto

Los vectores generados utilizan la polaridad de los términos para determinar la posición en el vector de características creado. Cabe dejar claro que dependiendo del modelo de datos los términos pueden ser unigramas, bigramas o skipgramas. En el caso de los unigramas, por ejemplo, si se construye un vector con la frecuencia de los términos según su polaridad con valores de polaridad desde -1.0 hasta 1.0, el vector que se obtiene sería como el que se muestra en la figura 2. En este vector por ejemplo se muestran dos términos con polaridad, según diccionario, entre el -0.8 y -0.9, un término con polaridad entre 0.1 y 0.2, y otro con polaridad mayor a 0.9. En este caso, en nuestro diccionario, la polaridad se representa con valores distribuidos desde lo más negativo hasta lo positivo con valores entre -1.0 y 0 para los negativos y 0 a 1.0 para los positivos.

Para el taller TASS2016 quisimos evaluar inicialmente el uso de vectores con la menor dimensión posible, así que en lugar de vectores de 20 celdas utilizamos solo vectores de 5 celdas para cada grupo de características, en lugar de saltos de 0.1 el rango utilizado es de

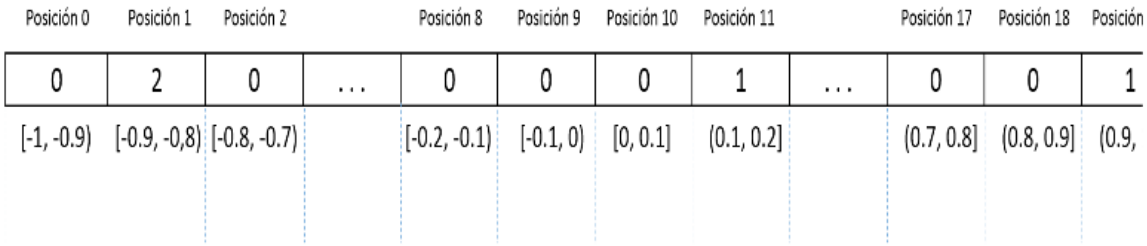

Figura 2: Vector de características

0.5.

3.4 Modelos locales y globales de representación del lenguaje

Nuestra propuesta pretende representar en los vectores de características información propia obtenida durante el proceso de entrenamiento, al igual que datos que representen información obtenida de modelos de lenguaje del español en general. En nuestro caso se utilizó inicialmente el diccionario generado a partir del corpus recolectado como insumo para obtener de él la información general del español. En el momento de entrenamiento, la polaridad de los términos en cada *tweet* son **conocidos** para ese conjunto de datos. La información global es la que se ha calculado previamente y se encuentra almacenada en forma de diccionarios. En nuestra propuesta lo que queremos hacer es representar en el vector las frecuencias de los términos de cada *tweet* distribuidos según su polaridad pero utilizar diferentes modelos de representación de lenguaje para llevar a cabo este cálculo. El diccionario utilizado en estos experimentos fue nuestra versión con unigramas. Se pretende utilizar representaciones con bigramas y una versión de skipgramas que incluye solo los términos anteriores a la palabra que se desea representar. Durante el entrenamiento, la polaridad obtenida en forma local es almacenada al igual que las frecuencias tomadas de diccionarios de polaridad global. Por lo tanto, los vectores cuentan con entradas para las distribuciones de polaridad local y las distribuciones de polaridad global. Aquí es donde incorporamos los diferentes modelos de lenguaje. Inicialmente trabajamos con **unigramas** para obtener resultados base para posteriores experimentos. Posteriormente, se genera un diccionario para **bigramas** y otro para lo que definimos como **skip-gramas previos**. Por el momento estas variantes no fueron enviadas como experimentos a TASS2016 sino solo las versiones iniciales.

4 Metodología

Utilizando el diccionario, el normalizador y el modelo de representación vectorial se procedió a crear vectores de respresentación con diferentes configuraciones. Primeramente se construyó una versión con vectores de dimensión 20 distribuyendo la polaridad de los términos según la polaridad almacenada para unigramas en el diccionario local. En este caso se pretende evaluar solamente el uso del diccionario y los marcadores de énfasis como repeticiones y mayúsculas. Este primer experimento es el denominado GASUCR-01. El segundo experimento consistió en evaluar un modelo un poco más robusto a nivel local con bigramas y la polaridad para el unigrama en el diccionario, si el bigrama no está presente durante el proceso de evaluación. En este caso se crearon vectores de menor dimensión para los datos locales, con solo cinco campos. Esta ejecución se idendificó como experimento GASUCR-01-noEMO-noPartNeg. Esta es la implementación base para luego evaluar el uso de bigramas tomados del contexto global. Esta versión base también fue enviada a la tarea de 4 categorías. En este caso, lo que se hizo fue unir las categorías +P y P en una sola, y la categoría +N con la N. El tercer experimento agregaba al anterior el uso de los emoticones, aparición de términos positivos con énfasis y las partículas negativas. En los resultados esta versión se identificó como GASUCR-04 En esta versión de TASS no nos dió tiempo de ejecutar las versiones con bigramas globales, ni skipgramas.

5 Resultados

Los resultados oficiales obtenidos para las ejecuciones antes mencionadas son los que se muestran en las Tablas 1 y 2. En estas figuras la columna **Ac.** muestra la *exactitud*, **P** se refiere a la **Macro Precisión**, **R** al **Macro Exhaustividad** y **F1** al **Macro F1**. En los resultados generales de TASS los resultados del grupo aparecen con el id indicado bajo el nombre del grupo GASUCR. En nuestro caso con el experimento 01 obtenemos los casos base para el uso de unigramas globales con vectores de dimensión 20 y los bigramas locales con dimensión 5. Es importante observar que los bigramas locales con dimensión 5 y las características de énfasis positivo, partículas de negación y emoticones producen un leve incremento pasando de 0.32 a 0.41. Otro aspecto que rescatamos es el aumento de la exactitud al pasar a la tarea de 3 categorías.

Tabla 1: Resultados Tarea 1 con 5 levels y corpus completo)

id	*Ac.*	*P*	*R*	*F1*
01	0.342	0.217	0.237	0.227
01-noEmNeg	0.326	0.334	0.258	0.291
04	0.410	0.268	0.242	0.254

Tabla 2: Resultados Tarea 1 con 3 niveles y corpus completo

id	*Ac.*	*P*	*R*	*F1*
01-noEmNeg	0.373	0.212	0.303	0.250

Estos casos se fueron seleccionando para ir evaluando en forma incremental cada uno de los aspectos relacionados a nuestra propuesta. Con cada característica nueva se trata de determinar su impacto sobre los valores de exactitud, precisión y exhaustividad.

6 Conclusiones y trabajo futuro

El marco de evaluación de TASS es provechoso para los grupos que inician la investigación en análisis de sentimiento en español con el fin de extenderla a otras latitudes. En nuestro caso pudimos evaluar y comparar la calidad de los resultados de los primeros casos base de nuestro trabajo. Observamos los primeros resultados con un sistema que utiliza un método de normalización con identificación de potenciales marcadores de énfasis, un modelo de representación basado en vectores de baja dimensión, y modelos de representación del texto con características locales y globales. El trabajo además hace uso de características comunes con otros como los son el uso de emoticones y partículas negativas. Como trabajo futuro tenemos pendiente la evaluación usando 3 categorías de los datos que hacen uso de contexto local con bigramas y características adicionales como uso de emoticones, palabras positivas con énfasis, y partículas de negación. Esperamos que los mejores resultados sean obtenidos al incorporar los nuevos modelos de lenguaje que estamos calculando para bigramas y skipgramas previos al unirlo con nuestro método de representación en vectores de baja dimensión. Se desea estudiar el efecto de la reducción del tamaño del vector al igual que técnicas de extrapolación de la polaridad en los modelos para los términos que no aparecen en los datos de entrenamiento.

Bibliografía

Batista, F. y R. Ribeiro. 2013. Sentiment analysis and topic classification based on binary maximum entropy classifiers. *Procesamiento de Lenguaje Natural*, 50:77–84.

Cabanlit, M. A. y K. Junshean Espinosa. 2014. Optimizing n-gram based text feature selection in sentiment analysis for commercial products in twitter through polarity lexicons. En *Information, Intelligence, Systems and Applications, IISA 2014, The 5th International Conference on*, páginas 94–97. IEEE.

Cambria, E., B. Schuller, Y. Xia, y C. Havasi. 2013. New avenues in opinion mining and sentiment analysis. *Intelligent Systems, IEEE*, PP(99):1–1.

Díaz-Galiano, M. y A. Montejo-Ráez. 2015. Participación de sinai dw2vec en tass 2015. En *Proceedings del Taller TASS 2015 en Análisis de Sentimiento de la XXXI Conferencia SEPLN 2015*, páginas 59–64.

Feldman, R. 2013. Techniques and applications for sentiment analysis. *Commun. ACM*, 56(4):82–89, Abril.

García-Cumbreras, M., J. Villena-Román, E. Martínez Cámara, M. C. Díaz-Galiano, M. T. Martín Valdivia, y L. A. Ureña López. 2016. Overview of

tass 2016. En *Proceedings of TASS 2016: Workshop on Sentiment Analysis at SEPLN co-located with the 32nd SEPLN Conference (SEPLN 2016)*, Salamanca, Spain, September.

Guo, L. y X. Wan. 2012. Exploiting syntactic and semantic relationships between terms for opinion retrieval. *Journal of the american society for information science and technology*, 63(11):2269–2282, Noviembre.

Indurkhya, N. y F. J. Damerau. 2010. *Handbook of natural language processing*, volumen 2. CRC Press.

Kiritchenko, S., X. Zhu, y S. M. Mohammad. 2014. Sentiment analysis of short informal texts. *Journal of Artificial Intelligence Research*, páginas 723–762.

Martínez-Cámara, E., M. Á. García-Cumbreras, M. T. Martín-Valdivia, y L. A. Ureña-L'opez. 2015. Sinai-emma: Vectores de palabras para el análisis de opiniones en twitter. En *Proceedings del Taller TASS 2015 en Análisis de Sentimiento de la XXXI Conferencia SEPLN 2015*, páginas 41–46.

Melero, M., A.-B. Cardús, A. Moreno, G. Rehm, K. de Smedt, y H. Uszkoreit. 2012. *The Spanish language in the digital age*. Springer.

Sharma, A. y S. Dey. 2012. A comparative study of feature selection and machine learning techniques for sentiment analysis. En *Proceedings of the 2012 ACM Research in Applied Computation Symposium*, páginas 1–7. ACM.

Turney, P. D. 2002. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. En *Proceedings of the 40th annual meeting on association for computational linguistics*, páginas 417–424. Association for Computational Linguistics.

Villena-Román, J., J. García Morera, M. Á. García-Cumbreras, E. M. Cámara, M. T. M. Valdivia, y L. A. U. López. 2015. Overview of tass 2015. En *Proceedings del Taller TASS 2015 en Análisis de Sentimiento de la XXXI Conferencia SEPLN 2015*, páginas 13–21.

Wang, X., F. Wei, X. Liu, M. Zhou, y M. Zhang. 2011. Topic sentiment analysis in twitter: a graph-based hashtag sentiment classification approach. En *Proceedings of the 20th ACM international conference on Information and knowledge management*, páginas 1031–1040. ACM.

TASS 2016: Workshop on Sentiment Analysis at SEPLN, septiembre 2016, pág. 29-33

LABDA at the 2016 TASS challenge task: using word embeddings for the sentiment analysis task*

LABDA en la competición TASS 2016: utilizando vectores de palabras para la tarea de análisis de sentimiento

Antonio Quirós[1,2], Isabel Segura-Bedmar[1], and Paloma Martínez[1]

[1]Departamento de Informática, Universidad Calos III de Madrid
Avd. de la Universidad, 30, 28911, Leganés, Madrid, España
100342879@alumnos.uc3m.es, isegura,pmf@inf.uc3m.es

[2]Sngular Data&Analytics
Av. LLano Castellano 13, Planta 5, 28034 Madrid, España
antonio.quiros@sngular.team

Resumen: Este artículo describe la participación del grupo LABDA en la tarea 1 (Sentiment Analysis at global level) de la competición TASS 2016. En nuestro enfoque, los tweets son representados por medio de vectores de palabras y son clasificados utilizando algoritmos como SVM y regresión logística.
Palabras clave: Análisis de Sentimiento, Vectores de palabras

Abstract: This paper describes the participation of the LABDA group at the Task 1 (Sentiment Analysis at global level). Our approach exploits word embedding representations for tweets and machine learning algorithms such as SVM and logistics regression.
Keywords: Sentiment Analysis, Word embeddings

1 Introduction

Knowing the opinion of customers or users has become a priority for companies and organizations in order to improve the quality of their services and products. With the ongoing explosion of social media, it affords a significant opportunity to poll the opinion of many Internet users by processing their comments. However, it should be noted that sentiment analysis, which can be defined as the automatic analysis of opinion in texts (Pang and Lee, 2008), is a challenging task because it is not strange that different people assign different polarities to a given text. On Twitter, the task is even more difficult, because the texts are small (only 140 characters) and are charectized by their informal style language, many grammatical errors and spelling mistakes, slang and vulgar vocabulary and abbreviations.

Since their introduction in 2013, the TASS shared task editions have had as main goal to promote the development of methods and resources for sentiment analysis of tweets in Spanish. This paper describes the participation of the LABDA group at the Task 1 (Sentiment Analysis at global level). In this task, the participating systems have to determine the global polarity of each tweet in the test dataset. There are two different evaluations: one based on 6 different polarity labels (P+, P, NEU, N, N+, NONE) and another based on just 4 labels (P, N, NEU, NONE). A detailed description of the task can be found in the overview paper of TASS 2016 (García-Cumbreras et al., 2016). Our approach exploits word embedding representations for tweets and machine learning algorithms such as SVM and logistics regression. The word embedding model can yield significant dimensionality reduction compared to the classical Bag-Of-Word (BoW) model. The dimensionality redution can have several positive effects on our algorithms such as faster training, avoiding overfitting and better performance.

The paper is organized as follows. Section 2 describes our approach. The experimental results are presented and discussed in Section

* This work was supported by eGovernAbility-Access project (TIN2014-52665-C2-2-R).

3. We conclude in Section 4 with a summary of our findings and some directions for future work.

2 System

In this paper, we study the use of word embeddings (also known as word vectors) in order to represent tweets and then examine several machine learning algorithms to classify them. Word embeddings have shown promising results in NLP tasks, such as named entity recognition (Segura-Bedmar, Suárez-Paniagua, and Martınez, 2015), relation extraction (Alam et al., 2016), sentiment analysis (Socher et al., 2013b) or parsing (Socher et al., 2013a). A word embedding is a function to map words to low dimensional vectors, which are learned from a large collection of texts. At present, Neural Network is one of the most used learning techniques for generating word embeddings (Mikolov and Dean, 2013). The essential assumption of this model is that semantically close words will have similar vectors (in terms of cosine similarity). Word embeddings can help to capture semantic and syntactic relationships of the corresponding words.

While the well-known Bag-of-Words (BoW) model involves a very large number of features (as many as the number of non-stopwords words with at least a minimum number of occurrences in the training data), the word embedding representation allows a significant reduction in the feature set size (in our case, from million to just 300). The dimensionality reduction is a desirable goal, because it helps in avoiding overfitting and leads to a reduction of the training and classification times, without any performance loss.

As a preprocessing step, tweets must be cleaned. First, we remove all links and urls. We then remove usernames which can be easily recognized because their first character is the symbol @. We then transform the hashtags to words by removing its first character (that is, the symbol #). Taking advantage of regular expressions, the emoticons are detected and classified in order to count the number of positive and negative emoticons in each tweet and then we remove them from the text. Table 1 shows the list of positive and negative emoticons, which were taken from the wikipedia page `https://en.wikipedia.org/wiki/List\_of\_emoticons`. We convert the tweets to lowercase and replace misspelled accented letters with the correct one (for instance "à" with "á"). We also treat elongations (that is, the repetition of a character) by removing the repetition of a character after its second occurrence (for example, "hoooolaaaa" would be translated to "hola"). We then decided to take into account laughs (for instance "jajaja") which turned out to be challenging because of the diverse ways they are expressed (i.e. expressions like "jajajaja" or "jejeje" and even misspelled ones like "jajjajaaj") We addressed this using regular expressions to standardize the different forms (i.e. "jajjjaaj" to "jajaja") and then replace them with the word "risas". Finally we remove all non-letters characters and all stopwords present in tweets[1].

Orientation	Emoticons
Positive	:-), :), :D, :o), :], D:3, :c), :>, =], 8), =), :}, :ˆ), :-D, 8-D, 8D, x-D, xD, X-D, XD, =-D, =D, =-3, =3, BˆD, :'), :'), :*, :-*, :ˆ*, ;-), ;), *-), *), ;-], ;], ;D, ;ˆ), >:P, :-P, :P, X-P, x-p, xp, XP, :-p, :p, =p, :-b, :b
Negative	>:[, :-(, :(, :-c, :-<, :<, :-[, :[, :{, ;(, :-||, >:(, :'-(, :'(, D:<, D=, v.v

Table 1: List of positive and negative emoticons

Once the tweets are preprocessed, they are tokenized using the NLKT toolkit (a Python package for NLP); we also performed experimentation by lemmatizing each tweet using MeaningCloud[2] Text Analytic software to compare both approaches. Then, for each token, we search its vector in the word embedding model. We use a pretrained model (Cardellino, 2016), which was generated by using the word2vec algorithm (Mikolov and Dean, 2013) from a collection of Spanish texts with approximately 1.5 billion words. The dimension of the word embedding is 300. It

[1] http://snowball.tartarus.org/algorithms/spanish/stop.txt
[2] `https://www.meaningcloud.com/`

should be noted that these texts were taken from different resources such as Spanish Wikipedia, WikiSource and Wikibooks, but none of them contains tweets. Therefore, it is possible that the main characteristics of the social media texts (such as informal style language, noisy, plenty of grammatical errors and spelling mistakes, slang and vulgar vocabulary, abbreviations, etc) are not correctly represented in this model. One of the main problems is that there is a significant number of words (almost a 13 % of the vocabulary, representing the 6 % of words occurrences) that are not found in the model. We perform a review of a small sample of these words, showing that most of them were mainly hashtags.

In our approach, a tweet of n tokens ($T = w_1, w_2, ..., w_n$) is represented as the centroid of the word vectors $\vec{w_i}$ of its tokens, as shown in the following equation:

$$\vec{T} = \frac{1}{n}\sum_{i=1}^{n}\vec{w_i} = \frac{\sum_{j=1}^{N}\vec{w_j}.TF(w_j,t)}{\sum_{j=1}^{N}TF(w_j,t)} \quad (1)$$

where N is the vocabulary size, that is, the total number of distinct words, while $TF(w_j, t)$ refers to the number of occurrences of the j-th vocabulary word in the tweet T.

We also explore the effect of including the inverse document frequencies IDF to represent tweets (see Equation 2). This helps to increase the weight of words that occur often, but only in a few documents, while it reduces the relevance of words that occur very frequently in a larger number of texts.

$$\vec{T} = \frac{1}{n}\sum_{i=1}^{n}\vec{w_i} = \frac{\sum_{j=1}^{N}\vec{w_j}.TF(w_j,t).IDF(w_j)}{\sum_{j=1}^{N}TF(w_j,t).IDF(w_j)} \quad (2)$$

having $IDF(w_j) = \frac{log|D|}{|tw \in D : w_j \in tw|}$ where $|D|$ refers to the number of tweets.

In addition to using the centroid, we assess the impact of complementing the tweet model with the following additional features:

- posWords: number of positive words present in the tweet.
- negWords: number of negative words present in the tweet.
- posEmo: number of positive emoticons present in the tweet.
- negEmo: number of negative emoticons present in the tweet.

For the posWords and negWords features we used the iSOL lexicon(Molina-González et al., 2013), a list composed by 2,509 positive words and 5,626 negative words. As described before, for the emoticons we used the listed in Table 1, but also added to the positive ones the number of laughs detected; and also, we included the number of recommendations present in the form of a "Follow Friday" hashtag (#FF), due to its ease of detection and its positive bias.

Classification is performed using scikit-learn, a Python module for machine learning. This package provides many algorithms such as Random Forest, Support Vector Machine (SVM) and so on. One of its main advantages is that it is supported by extensive documentation. Moreover, it is robust, fast and easy to use.

As stated before, we have two main training models: Averaged centroids and the averaged centroids including the inverted document frequency, for both the lemmatized and not-lemmatized texts. We performed experiments using three different classifiers: Random Forests, Support Vector Machines and Logistic Regression because these classifiers often achieved the best results for text classification and sentiment analysis.

Also we evaluated the impact of applying a set of emoticon's rules as a pre-classification stage, similar to (Chikersal et al., 2015), in which we determine a first stage polarity for each tweet as follows:

- If posEmo is greater than zero and negEmo is equal to zero, the tweet is marked as "P".
- If negEmo is greater than zero and posEmo is equal to zero, the tweet is marked as "N".
- If both posEmo and negEmo are greater than zero, the tweet is marked as "NEU".
- If both posEmo and negEmo are equal to zero, the tweet is marked as "NONE".

Then, after the classification takes place we made three tests: i) Applying no rule, ii) honoring the polarity defined by the rule, which means, we keep the predefined polarity

if the tweet was marked as "P" or "N", otherwise we take the value estimated by the classifier, and iii) a mixed approach where we give each polarity a value (N+: -2; N: -1; NEU,NONE: 0; P: 1; P+: 2) and performed an arithmetic sum of both the predefined and estimated polarity if and only if they are not equal; with that for instance, if the classifier marked a tweet as "N" and the rules marked it as "P" the tweet will be classified as "NEU".

3 Results

In order to choose the best-performing classifiers, we use 10-fold cross-validation because there is no development dataset and this strategy has become the standard method in practical terms. Our experiments showed that, although the results were similar[3], the best settings for the 5-levels task are:

- RUN-1: Support Vector Machine, over the averaged centroids without applying any rules for pre-defining polarities.
- RUN-2: Support Vector Machine, over the averaged centroids and applying the mixed rules approach.
- RUN-3: Logistic Regression, over the centroids with inverted document frequency and applying the mixed rules approach.

and for the 3-levels task are:

- RUN-1: Support Vector Machine, over the averaged centroids and applying the mixed rules approach.
- RUN-2: Logistic Regression, over the centroids with inverted document frequency and applying the mixed rules approach.
- RUN-3: Logistic Regression, over the averaged centroids and applying the mixed rules approach.

Tables 2 and 3 show the results for these settings provided by the TASS submission system. For each run, accuracy is provided as well as the macro-averaged precision, recall and F1-measure. As expected, the results for 3 levels are higher than for 5 levels because the training dataset is larger.

[3]Experiments showed that not-lemmatized text performed better in all settings, hence the best settings reported here is using not-lematized model

Run	P	R	F1	Acc
RUN-1	0.411	0.449	0.429	0.527
RUN-2	0.412	0.448	0.429	0.527
RUN-3	0.402	0.436	0.418	0.549

Table 2: Results for Sentiment Analysis at global level (5 levels, Full test corpus)

Run	P	R	F1	Acc
RUN-1	0.506	0.510	0.508	0.652
RUN-2	0.508	0.508	0.508	0.652
RUN-3	0.512	0.511	0.511	0.653

Table 3: Results for Sentiment Analysis at global level (3 levels, Full test corpus)

With the settings mentioned above, the obtained results are extremely similar, but we can state that, in terms of Accuracy, Logistic Regression report the best results; and, even it's not measured in this work, is worth mentioning that Logistic Regression's performance was observably faster.

4 Conclusions and future work

This paper explores the use of word embeddings for the task of sentiment analysis. Instead of using, the bag-of-words model to represent tweets, these are represented as word vectors taken from a pre-trained model of word embeddings. An important advantage of word embedding model compared to the technique of bag-of-words representation is that it achieves a significant dimensional reduction of the feature set needed to represent tweets and leads, therefore, to a reduction of training and testing time of the algorithms.

In order to use word embedding models properly, a preprocessing stage had to be completed before training a classifier. Due to the unstructured nature of the tweets, this preprocessing proved to be a very important step in order to standardize at some degree the input data. The experimentation showed that the three tested classifiers obtained very similar results, with Random Forest having slight worse performance and Logistic Regression being slightly better and much more faster.

One of the main drawback of our approach is that many words do not have a word vector in the word embedding model used for our experiments. An analysis showed that many

of these words come from hashtags, which are usually short phrases. Therefore, we should apply a more sophisticated method in order to extract the words forming hashtag.

As future work, we also plan to use a word embedding model trained on a collection of text from Spanish social media. We think that this will have a positive effect of the performance of our system to identify the polarity of tweets because this model will be generated from documents characterized by the main features that describe social media texts (for example, informal style language, plenty of grammatical errors and spelling mistakes, slang and vulgar vocabulary).

Acknowledgments

This work was supported by eGovernAbility-Access project (TIN2014-52665-C2-2-R).

References

Alam, F., A. Corazza, A. Lavelli, and R. Zanoli. 2016. A knowledge-poor approach to chemical-disease relation extraction. *Database*, 2016:baw071.

Cardellino, C. 2016. Spanish Billion Words Corpus and Embeddings, March.

Chikersal, P., S. Poria, E. Cambria, A. Gelbukh, and C. E. Siong. 2015. Modelling public sentiment in twitter: using linguistic patterns to enhance supervised learning. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 49–65. Springer.

García-Cumbreras, M. A., J. Villena-Román, E. Martínez-Cámara, M. C. Díaz-Galiano, M. T. Martín-Valdivia, and L. A. U. na López. 2016. Overview of tass 2016. In *Proceedings of TASS 2016: Workshop on Sentiment Analysis at SEPLN colocated with the 32nd SEPLN Conference (SEPLN 2016)*, Salamanca, Spain, September.

Mikolov, T. and J. Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*.

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.

Pang, B. and L. Lee. 2008. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2):1–135.

Segura-Bedmar, I., V. Suárez-Paniagua, and P. Martınez. 2015. Exploring word embedding for drug name recognition. In *SIXTH INTERNATIONAL WORKSHOP ON HEALTH TEXT MINING AND INFORMATION ANALYSIS (LOUHI)*, page 64.

Socher, R., J. Bauer, C. D. Manning, and A. Y. Ng. 2013a. Parsing with compositional vector grammars. In *ACL (1)*, pages 455–465.

Socher, R., A. Perelygin, J. Y. Wu, J. Chuang, C. D. Manning, A. Y. Ng, and C. Potts. 2013b. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the conference on empirical methods in natural language processing (EMNLP)*, volume 1631, page 1642. Citeseer.

TASS 2016: Workshop on Sentiment Analysis at SEPLN, septiembre 2016, pág. 35-39

JACERONG at TASS 2016: An Ensemble Classifier for Sentiment Analysis of Spanish Tweets at Global Level

JACERONG en TASS 2016: Combinación de clasificadores para el análisis de sentimientos de tuits en español a nivel global

Jhon Adrián Cerón-Guzmán
Santiago de Cali, Valle del Cauca, Colombia
jadrian.ceron@gmail.com

Resumen: Este artículo describe un enfoque basado en conjuntos de clasificadores que se ha desarrollado para participar en la Tarea 1 del taller TASS sobre análisis de sentimientos de tuits en español a nivel global. Los conjuntos se construyen sobre la combinación de sistemas con la correlación absoluta más baja entre sí. Estos sistemas son capaces de tratar con formas léxicas no estándar en los tweets, con el fin de mejorar la calidad del análisis de lenguaje natural. Para realizar la clasificación de polaridad, el enfoque utiliza características básicas que han probado su poder discriminativo, así como características de n-gramas de palabras y caracteres. Luego, las salidas de clasificadores de Regresión logística, que pueden ser etiquetas de clase o probabilidades para cada clase, se utilizan para construir conjuntos de clasificadores. Los resultados experimentales muestran que la combinación menos correlacionada de 25 sistemas, la cual elige la clase con la probabilidad promedio no poderada más alta, es la configuración que mejor se adapta a la tarea, alcanzando una precisión global de 62.0% en la evaluación de seis etiquetas, y de 70.5% en la evaluación de cuatro etiquetas.
Palabras clave: Análisis de sentimientos, clasificación de polaridad, combinación de clasificadores, normalización léxica, tuis en español, Twitter

Abstract: This paper describes an ensemble-based approach developed to participate in TASS-2016 Task 1 on sentiment analysis of Spanish tweets at global level. Ensembles are built on the combination of systems with the lowest absolute correlation with each other. The systems are able to deal with non-standard lexical forms in tweets, in order to improve the quality of natural language analysis. To support the polarity classification, the approach uses basic features that have proved their discriminative power, as well as word and character n-gram features. Then, outputs from Logistic Regression classifiers, which may be either class labels or probabilities for each class, are used to build ensembles. Experimental results show that the less-correlated combination of 25 systems, which chooses the class with the highest unweighted average probability, is the setting that best suits to the task, achieving an overall accuracy of 62.0% in the six-labels evaluation, and of 70.5% in the four-labels evaluation.
Keywords: Ensemble classifier, lexical normalization, polarity classification, sentiment analysis, Spanish tweets, Twitter

1 Introduction

What people say on social media about issues of their everyday life, the society, and the world in general, has turned into a rich source of information to understand social behavior. Twitter content, in particular, has caught the attention of researchers who have investigated its potential for conducting studies on the human subjectivity at large scale, which was not feasible using traditional methods. Around election time, sentiment analysis of political tweets has been widely used to capture trends in public opinion regarding important issues such as voting intention (Gayo-Avello, 2013). However, analyzing this content also presents several challenges, including the development of text analysis approaches based on Natural Language Processing techniques, which properly adapt to the informal genre and the free writ-

ISSN 1613-0073

ing style of Twitter (Han and Baldwin, 2011; Cerón-Guzmán and León-Guzmán, 2016).

TASS is a workshop aimed at fostering research on sentiment analysis of Spanish Twitter data, which provides a benchmark evaluation to compare the latest advances in the field (García-Cumbreras et al., 2016). One of the proposed tasks is to determine the opinion orientation expressed in tweets at global level. Task 1 consists on assigning one of six labels (P+, P, NEU, N, N+, NONE) to a tweet in the six-labels evaluation; or one of four labels (P, NEU, N, NONE) in the four-labels evaluation. Here, P, N, and NEU, stand for positive, negative, and neutral, respectively; NONE, instead, means no sentiment. The "+" symbol is used as intensifier.

This paper presents an ensemble-based approach to polarity classification of Spanish tweets, developed to participate in Task 1 proposed by the organizing committee of the TASS workshop. The ensemble members are (relatively) highly correct classifiers with the lowest absolute correlation with each other. The output from each classifier, which may be either a class label or probabilities for each class, is used to assign the polarity to a tweet based on a majority rule or on the highest unweighted average probability. Moreover, classifiers are adapted to deal with non-standard lexical forms in tweets, in order to improve the quality of natural language analysis.

The remainder of this paper is organized as follows. Section 2 describes the common architecture of the ensemble members (i.e., classifiers). Next, the submitted experiments, as well as the obtained results, are discussed in Section 3. Finally, Section 4 concludes the paper.

2 The System Architecture

The tweet text is passed through the pipeline of each system in order to assign it a class label or a probability to be of a certain class. The pipeline, which goes from text preprocessing to machine learning classification, is described below. Note that the system term is preferred over the classifier term, because a machine learning classifier receives a feature vector and produces a class label or probabilities for each class; instead, the system term enables to conceive the whole process, from preprocessing to machine learning classification.

2.1 Preprocessing

The process of text cleaning and normalization is performed in two phases: basic preprocessing and advanced preprocessing.

2.1.1 Basic Preprocessing

The following simple rules are implemented as regular expressions:

- Removing URLs and emails.
- HTML entities are mapped to textual representations (e.g., "<" → "<").
- Specific Twitter terms such as mentions (@user) and hashtags (#topic) are replaced by placeholders.
- Unknown characters are mapped to their closest ASCII variant, using the Python *Unidecode* module for the mapping.
- Consecutive repetitions of a same character are reduced to one occurrence.
- Emoticons are recognized and then classified into positive and negative, according to the sentiment they convey (e.g., ":)" → "EMO_POS", ":(" → "EMO_NEG").
- Unification of punctuation marks (Vilares, Alonso, and Gómez-Rodrıguez, 2014).

2.1.2 Advanced Preprocessing

Once the set of simple rules has been applied, the tweet text is tokenized and morphologically analyzed by FreeLing (Padró and Stanilovsky, 2012). In this way, for each resulting token, its lemma and Part-of-Speech (POS) tag are assigned. Taking these data as input, the following advanced preprocessing is applied:

- **Lexical normalization.** Each token is passed through a set of basic modules of FreeLing (e.g., dictionary lookup, suffixes check, detection of numbers and dates, and named entity recognition) for identifying standard word forms and other valid constructions. If a token is not recognized by any of the modules, it is marked as out-of-vocabulary (OOV) word. Then, a confusion set is formed by normalization candidates which are identical or similar to the graphemes or phonemes that make the

OOV word. These candidates are elements of the union of a dictionary of Spanish standard word forms and a gazetteer of proper nouns. The best normalization candidate for the OOV word is which best fits a statistical language model. The language model was estimated from the Spanish Wikipedia corpus. Lastly, the selected candidate is capitalized according to the capitalization rules of the Spanish language. Extensive research on lexical normalization of Spanish tweets can be read in (Cerón-Guzmán and León-Guzmán, 2016).

- **Negation handling.** Inspired by the approach proposed by Pang et al. (Pang, Lee, and Vaithyanathan, 2002), this research defined a negated context as a segment of the tweet that starts with a (Spanish) negation word and ends with a punctuation mark (i.e., "¡", ",", ":", "?", ".", ";"), but only the first $n \epsilon [0, 3]$ or all tokens labeled with any or a specific POS tag (i.e., verb, adjective, adverb, and common noun) are affected by adding it the "_NEG" suffix. Note that when $n = 0$, no token is affected.

2.2 Feature Extraction

In this stage, the normalized tweet text is transformed into a feature vector that feeds the machine learning classifier. The features are grouped into basic features and n-gram features.

2.2.1 Basic Features

Some of these features are computed before the process of text cleaning and normalization is performed.

- The number of words completely in uppercase.
- The number of words with more than two consecutive repetitions of a same character.
- The number of consecutive repetitions of exclamation marks, question marks, and both punctuation marks (e.g., "¡!", "??", "?!") and whether the text ends with an exclamation or question mark.
- The number of occurrences of each class of emoticons (i.e., positive and negative) and whether the last token of the tweet is an emoticon.
- The number of positive and negative words, relative to the ElhPolar lexicon (Saralegi and Vicente, 2013), the AFINN lexicon (Nielsen, 2011), or an union of both lexicons. In a negated context, the label of a polarity word is inverted (i.e., positive words become negative words, and vice versa). Additionally, a third feature labels the tweet with the class whose number of polarity words in the text is the highest.
- The number of negated contexts.
- The number of occurrences of each Part-of-Speech tag.

2.2.2 N-gram Features

The fixed-length set of basic features is always extracted from tweets. However, the tweet text varies from another in terms of length, number of tokens, and vocabulary used. For that reason, a process that transforms textual data into numerical feature vectors of fixed length is required. This process, known as vectorization, is performed by applying the tf-idf weighting scheme (Manning, Raghavan, and Schütze, 2008). Thus, each document (i.e., a tweet text) is represented as a vector $d = \{t_1, \ldots, t_n\} \epsilon \mathbb{R}^V$, where V is the size of the vocabulary that was built by considering word n-grams with $n \epsilon [1, 4]$, or character n-grams with $n \epsilon [3, 5]$ in the collection (i.e., the training set). The vector is, hence, formed by word n-grams, character n-grams, or a concatenation of word and character n-grams.

2.3 Machine Learning Classification

At the last stage, the sentiment analysis system classifies a given tweet as either P+, P, NEU, N, N+, or NONE, or assigns probabilities for each class. After receiving as input the feature vector, a L2-regularized Logistic Regression classifier assigns a class label to the tweet or a probability to be of a certain class. The classifier was trained on the training set, using the Scikit-learn (Pedregosa et al., 2011) implementation of the Logistic Regression algorithm.

3 Experiments

1,720 different sentiment analysis systems were trained on the training set via 5-fold cross validation, in order to find the best parameter settings, namely: negation handling,

polarity lexicon, order of word and character n-grams, and others parameters related to the vectorization process (e.g., lowercasing, frequency thresholds, etc.). The systems were sorted by their mean cross-validation score, and thus the top 50 ranked were filtered to build the ensemble. The training set is a collection of 7,219 tweets, each of which is tagged with one of six labels (i.e., P+, P, NEU, N, N+, and NONE). Note that the systems were trained for the six-labels evaluation, and therefore the P+ and P labels were merged into P, as well as the N+ and N labels were merged into N, to produce an output in accordance with the four-labels evaluation. Further description of the provided corpus, as well as of the training and test sets, can be read in (García-Cumbreras et al., 2016).

Next, the top 50 systems assigned a class label to each tweet in a collection of 1,000, which was drawn from the untagged test set with a similar class distribution to the training set. In this stage, the objective was to find the systems with the lowest absolute correlation with each other; therefore, the performance was not evaluated. Then, the less-correlated combinations of 5, 10, and 25 systems, were used to build the ensembles, whose outputs correspond to the submitted experiments. These experiments are described below:

- **run-1**: the less-correlated combination of 5 systems, which chooses the class label that represents the majority in the predictions made by the ensemble members.
- **run-2**: the less-correlated combination of 10 systems, which chooses the class with the highest unweighted average probability.
- **run-3**: the less-correlated combination of 25 systems, which chooses the class with the highest unweighted average probability.

Tables 1 and 2 show the performance evaluation on the test set (i.e., a collection of 60,798 tweets) for six and four labels, respectively. Accuracy has been defined as the official metric for ranking the systems. In summary, the main gain occurs among the "run-1" and "run-2" experiments, with an increment of 0.5% in accuracy in the six-labels evaluation, and of 0.2% in the four-labels evaluation; instead, a negligible gain occurs among the "run-2" and" run-3" experiments, taking additionally into account the computational cost of running the latter.

Experiment	Accuracy	Macro-Precision	Macro-Recall	Macro-F1
run-1	0.614	0.471	0.531	0.499
run-2	0.619	0.476	0.535	0.504
run-3	0.620	0.477	0.532	0.503

Table 1: Performance on the test set in the six-labels evaluation

Experiment	Accuracy	Macro-Precision	Macro-Recall	Macro-F1
run-1	0.702	0.564	0.565	0.564
run-2	0.704	0.567	0.568	0.567
run-3	0.705	0.568	0.567	0.568

Table 2: Performance on the test set in the four-labels evaluation

Class	Precision	Recall	F1-score
P	0.755	0.786	0.770
NEU	0.128	0.093	0.107
N	0.631	0.812	0.710
NONE	0.758	0.578	0.656

Table 3: Discriminative power for each class in the four-labels evaluation

As a final point, Table 3 shows how the overall performance is affected by the low discriminative power of the ensembles (in this case, the one that correspond to "run-3") for the NEU class. With this in mind, it is proposed as future work to deal with the low representativeness of the NEU class in the training data (i.e., 9.28% of tweets), in order to properly characterize this kind of tweets.

4 Conclusion

This paper has described an ensemble-based approach for sentiment analysis of Spanish Twitter data at global level, developed in order to participate in Task 1 proposed by the organization of TASS workshop. Three ensembles were built on the combination of sentiment analysis systems with the lowest absolute correlation with each other. The systems were adapted to the informal genre and the free writing style that characterize Twitter, in order to improve the quality of natural language analysis. In this way, the predicted class label for a particular tweet

was based on a majority rule or on the highest average probability. Experimental results showed that the less-correlated combination of 25 systems, which chose the class with the highest unweighted average probability, was the setting that best suited to the task. However, there is a great room for improvement in the learning of a proper characterization of neutral tweets.

References

Cerón-Guzmán, J. A. and E. León-Guzmán. 2016. Lexical normalization of Spanish tweets. In *Proceedings of the 25th International Conference Companion on World Wide Web*, WWW'16 Companion, pages 605–610. International World Wide Web Conferences Steering Committee.

García-Cumbreras, M. A., J. Villena-Román, E. Martínez-Cámara, M. C. Díaz-Galiano, M. T. Martín-Valdivia, and L. A. Urena-López. 2016. Overview of tass 2016. In *Proceedings of TASS 2016: Workshop on Sentiment Analysis at SEPLN co-located with the 32nd SEPLN Conference (SEPLN 2016)*, Salamanca, Spain, September.

Gayo-Avello, D. 2013. A meta-analysis of state-of-the-art electoral prediction from Twitter data. *Soc. Sci. Comput. Rev.*, 31(6):649–679.

Han, B. and T. Baldwin. 2011. Lexical normalisation of short text messages: Makn sens a #Twitter. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT'11, pages 368–378, Stroudsburg, PA, USA. Association for Computational Linguistics.

Manning, C. D., P. Raghavan, and H. Schütze. 2008. Scoring, term weighting and the vector space model. In *An Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA.

Nielsen, F. Å. 2011. A new anew: evaluation of a word list for sentiment analysis in microblogs. In *Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts': Big things come in small packages*, pages 93–98.

Padró, L. and E. Stanilovsky. 2012. Freeling 3.0: Towards wider multilinguality. In *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)*, Istanbul, Turkey, May. ELRA.

Pang, B., L. Lee, and S. Vaithyanathan. 2002. Thumbs up?: Sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10*, EMNLP '02, pages 79–86. Association for Computational Linguistics.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Saralegi, X. and I. S. Vicente. 2013. Elhuyar at tass 2013. In *Proceedings of the Sentiment Analysis Workshop at SEPLN (TASS2013)*, September.

Vilares, D., M. A. Alonso, and C. Gómez-Rodrıguez. 2014. On the usefulness of lexical and syntactic processing in polarity classification of twitter messages. *Journal of the Association for Information Science and Technology*.

TASS 2016: Workshop on Sentiment Analysis at SEPLN, septiembre 2016, pág. 41-45

Participación de SINAI en TASS 2016*

SINAI participation in TASS 2016

A. Montejo-Ráez
University of Jaén
23071 Jaén (Spain)
amontejo@ujaen.es

M.C. Díaz-Galiano
University of Jaén
23071 Jaén (Spain)
mcdiaz@ujaen.es

Resumen: Este artículo describe el sistema de clasificación de la polaridad utilizado por el equipo SINAI en la tarea 1 del taller TASS 2016. Como en participaciones anteriores, nuestro sistema se basa en un método supervisado con SVM a partir de vectores de palabras. Dichos vectores se calculan utilizando la técnicas de *deep-learning* Word2Vec, usando modelos generados a partir de una colección de tweets expresamente generada para esta tarea y el volcado de la Wikipedia en español. Nuestros experimentos muestran que el uso de colecciones de datos masivos de Twitter pueden ayudar a mejorar sensiblemente el rendimiento del clasificador.
Palabras clave: Análisis de sentimientos, clasificación de la polaridad, deep-learning, Word2Vec

Abstract: This paper introduces the polarity classification system used by the SINAI team for the task 1 at the TASS 2016 workshop. Our approach is based on a supervised learning algorithm over vectors resulting from a weighted vector. This vector is computed using a deep-learning algorithm called Word2Vec. The algorithm is applied so as to generate a word vector from a deep neural net trained over a specific tweets collection and the Spanish Wikipedia. Our experiments show massive data from Twitter can lead to a slight improvement in classificaciones accuracy.
Keywords: Sentiment analysis, polarity classification, deep learning, Word2Vec, Doc2Vec

1 Introducción

En este trabajo describimos las aportaciones realizadas para participar en la tarea 1 del taller TASS (Sentiment Analysis at global level), en su edición de 2016 (García-Cumbreras et al., 2016). Nuestra solución continúa con las técnicas aplicadas en el TASS 2014 (Montejo-Ráez, García-Cumbreras, y Díaz-Galiano, 2014) y 2015 (Díaz-Galiano y Montejo-Ráez, 2015), utilizando aprendizaje profundo para representar el texto y una colección de entrenamiento creada con tweets que contienen emoticonos que expresan emociones de felicidad o tristeza. Para ello utilizamos el método *Word2Vec*, ya que ha obtenido los mejores resultados en años anteriores. Por lo tanto, generamos un vector de pesos para cada palabra del tweet utilizando Word2Vec, y realizamos la media de dichos vectores para obtener una única representación vectorial. Nuestros resultados demuestran que el rendimiento del sistema de clasificación puede verse sensiblemente mejorado gracias a la introducción de estos datos en la generación del modelo de palabras, no así en el entrenamiento del clasificador de polaridad final.

La tarea del TASS en 2016 denominada *Sentiment Analysis at global level* consiste en el desarrollo y evaluación de sistemas que determinan la polaridad global de cada tweet del corpus general. Los sistemas presentados deben predecir la polaridad de cada tweet utilizando 6 o 4 etiquetas de clase (granularidad fina y gruesa respectivamente).

El resto del artículo está organizado de la siguiente forma. El apartado 2 describe el estado del arte de los sistemas de clasificación de polaridad en español. A continuación, se describe la colección de tweets con emoticonos utilizada para entrenar el clasificador. En el apartado 4 se describe el sistema desarro-

* Este estudio está parcialmente financiado por el proyecto TIN2015-65136-C2-1-R otorgado por el Ministerio de Economía y Competitividad del Gobierno de España.

llado y en el apartado 5 los experimentos realizados, los resultados obtenidos y el análisis de los mismos. Finalmente, en el último apartado exponemos las conclusiones y el trabajo futuro.

2 Clasificación de la polaridad en español

La mayor parte de los sistemas de clasificación de polaridad están centrados en textos en inglés, y para textos en español el sistema más completo, en cuanto a técnicas lingüísticas aplicadas, posiblemente sea *The Spanish SO Calculator* (Brooke, Tofiloski, y Taboada, 2009), que además de resolver la polaridad de los componentes clásicos (adjetivos, sustantivos, verbos y adverbios) trabaja con modificadores como la detección de negación o los intensificadores.

Los algoritmos de aprendizaje profundo (*deep-learning* en inglés) están dando buenos resultados en tareas donde el estado del arte parecía haberse estancado (Bengio, 2009). Estas técnicas también son de aplicación en el procesamiento del lenguaje natural (Collobert y Weston, 2008), e incluso ya existen sistemas orientados al análisis de sentimientos, como el de Socher et al. (Socher et al., 2011). Los algoritmos de aprendizaje automático no son nuevos, pero sí están resurgiendo gracias a una mejora de las técnicas y la disposición de grandes volúmenes de datos necesarios para su entrenamiento efectivo.

En la edición de TASS en 2012 el equipo que obtuvo mejores resultados (Saralegi Urizar y San Vicente Roncal, 2012) presentaron un sistema completo de pre-procesamiento de los tweets y aplicaron un lexicón derivado del inglés para polarizar los tweets. Sus resultados eran robustos en granularidad fina (65 % de accuracy) y gruesa (71 % de accuracy).

En la edición de TASS en 2013 el mejor equipo (Fernández et al., 2013) tuvo todos sus experimentos en el top 10 de los resultados, y la combinación de ellos alcanzó la primera posición. Presentaron un sistema con dos variantes: una versión modificada del algoritmo de ranking (RA-SR) utilizando bigramas, y una nueva propuesta basada en skipgrams. Con estas dos variantes crearon lexicones sobre sentimientos, y los utilizaron junto con aprendizaje automático (SVM) para detectar la polaridad de los tweets.

En 2014 el equipo con mejores resultados en TASS se denominaba ELiRF-UPV (Hurtado y Pla, 2014). Abordaron la tarea como un problema de clasificación, utilizando SVM. Utilizaron una estrategia uno-contra-todos donde entrenan un sistema binario para cada polaridad. Los tweets fueron tokenizados para utilizar las palabras o los lemas como características y el valor de cada característica era su coeficiente tf-idf. Posteriormente realizaron una validación cruzada para determinar el mejor conjunto de características y parámetros a utilizar.

El equipo ELiRF-UPV (Hurtado, Pla, y Buscaldi, 2015) volvió a obtener los mejores resultados en la edición de TASS 2015 con una técnica muy similar a la edición anterior (SVM, tokenización, clasificadores binarios y coeficientes tf-idf). En este caso utilizaron un sistema de votación simple entre un mayor número de clasificadores con parámetros distintos. Los mejores resultados los obtuvieron con un sistema que combinaba 192 sistemas SVM con configuraciones diferentes, utilizando un nuevo sistema SVM para realizar dicha combinación.

3 Colección de tweets con emoticonos

Los algoritmos de deep-learning necesitan grandes volúmenes de datos para su entrenamiento. Por ese motivo se ha creado una colección de tweets específica para la detección de polaridad. Para crear dicha colección se han recuperado tweets con las siguientes características:

- Que contengan emoticonos que expresen la polaridad del tweet. En este caso se han utilizado los siguientes emoticonos:
  - Positivos: `:)` `:-)` `:D` `:-D`
  - Negativos: `:(` `:-(`
- Que los tweets no contengan URLs, para evitar tweets cuyo contenido principal se encuentra en el enlace.
- Que no sean retweets, para reducir el número de tweets repetidos.

La captura de dichos tweets se realizó durante 22 días, del 18/07/2016 hasta el 9/08/2016, recuperando unos 100.000 tweets diarios aproximadamente. Tal y como se ve en la Figura 1 la recuperación fue muy homogénea y se obtuvieron más de 2.000.000 de tweets.

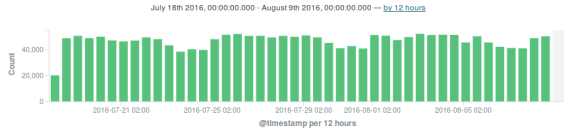

Figura 1: Número de tweets recuperados cada 12 horas

Posteriormente, se realizó un filtrado de dichos tweets eliminando aquellos que contubieran menos de 5 palabras, teniendo en cuenta que consideramos palabra todo término que sólo contenga letras (sin números, ni caracteres especiales).

Al final quedaron 1.777.279 clasificados según el emoticono que contienen de la siguiente manera:

- Positivos: 869.339 tweets
- Negativos: 907.940 tweets

Por último, se realiza la siguiente limpieza de tweets:

- Convertir el texto a minúsculas.
- Eliminar menciones (nombres de usuario que empiezan el caracter @).
- Sustituir letras acentuadas por sus versiones sin acentuar.
- Quitar las palabras vacías de contenido (*stopwords*).
- Normalizar las palabras para que no contengan letras repetidas, sustituyendo las repeticiones de letras contiguas para dejar sólo 3 repeticiones.

4 *Descripción del sistema*

Word2Vec[1] es una implementación de la arquitectura de representación de las palabras mediante vectores en el espacio continuo, basada en bolsas de palabras o n-gramas concebida por Tomas Mikolov et al. (Mikolov et al., 2013). Su capacidad para capturar la semántica de las palabras queda comprobada en su aplicabilidad a problemas como la analogía entre términos o el agrupamiento de palabras. El método consiste en proyectar las palabras a un espacio n-dimensional, cuyos pesos se determinan a partir de una estructura de red neuronal mediante un algoritmo recurrente. El modelo se puede configurar para que utilice una topología de bolsa de palabras (CBOW) o *skip-gram*, muy similar al anterior, pero en la que se intenta predecir los términos acompañantes a partir de un término dado. Con estas topologías, si disponemos de un volumen de textos suficiente, esta representación puede llegar a capturar la semántica de cada palabra. El número de dimensiones (longitud de los vectores de cada palabra) puede elegirse libremente. Para el cálculo del modelo Word2Vec hemos recurrido al software indicado, creado por los propios autores del método.

Tal y como se ha indicado, para obtener los vectores Word2Vec representativos para cada palabra tenemos que generar un modelo a partir de un volumen de texto grande. Para ello hemos utilizado los parámetros que mejores resultados obtuvieron en nuestra participación del 2014 (Montejo-Ráez, García-Cumbreras, y Díaz-Galiano, 2014). Por lo tanto, a partir de un volcado de Wikipedia[2] en Español de los artículos en XML, hemos extraído el texto de los mismos. Obtenemos así unos 2,2 GB de texto plano que alimenta al programa *word2vec* con los parámetros siguientes: una ventana de 5 términos, el modelo *skip-gram* y un número de dimensiones esperado de 300, logrando un modelo con más de 1,2 millones de palabras en su vocabulario.

Como puede verse en la Figura 2, nuestro sistema realiza la clasificación de los tweets utilizando dos fases de aprendizaje, una en la que entrenamos el modelo Word2Vec haciendo uso de un volcado de la enciclopedia on-line Wikipedia, en su versión en español, como hemos indicado anteriormente. De esta forma representamos cada tweet con el vector resultado de calcular la media de los vectores Word2Vec de cada palabra en el tweet y su desviación típica (por lo que cada vector de palabras por modelo es de 600 dimensiones). Se lleva a cabo una simple normalización previa sobre el tweet, eliminando repetición de letras y poniendo todo a minúsculas. La segunda fase de entrenamiento utiliza el algoritmo SVM y se entrena con la colección de tweets con emoticonos explicada en el apartado 3. La implementación de SVM utilizada es la basada en kernel lineal con entrenamiento SGD (Stochastic Gradient Descent) proporcionada por la biblioteca Sci-kit Learn[3] (Pedregosa et al., 2011).

Esta solución es la utilizada en las dos variantes de la tarea 1 del TASS con predicción

[1] `https://code.google.com/p/word2vec/`

[2] `http://dumps.wikimedia.org/eswiki`

[3] `http://scikit-learn.org/`

de 4 clases: la que utiliza el corpus de tweets completo (full test corpus) y el que utiliza el corpus balanceado (1k test corpus).

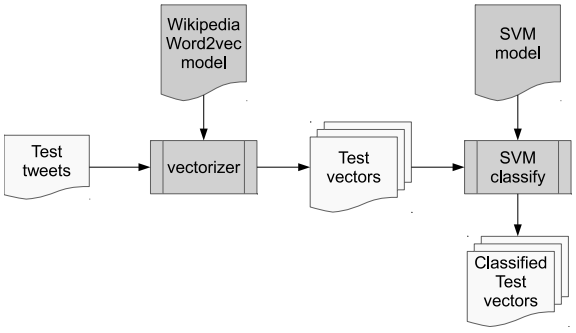

Figura 2: Flujo de datos del sistema completo

5 Resultados obtenidos

Hemos experimentado con el efecto que tienen en el rendimiento del sistema el uso de una colección de datos generada a partir de la captura de tweets y que han sido etiquetados según los emoticonos que contienen en la forma comentada anteriormente. La colección de más de 1,7 millones de tweets ha sido utilizada al completo para generar un modelo de vectores de palabras, cuya combinación con el de Wikipedia se ha analizado. También hemos comprobado cómo el uso de dicha colección de tweets afecta cuando se usa para el entrenamiento del modelo de clasificación de la polaridad. Para ello se han seleccionado 500,000 tweets aleatoriamente de esta colección, con sus correspondientes etiquetas P (positivo) o N (negativo) y se han combiando con la colecciónd de entrenamiento de TASS.

Los resultados según las medidaas de *Accuracy* y *Macro F1* obtenidas se muestran en la tabla 1. La primera columna nos indica a partir de cuáles datos se han generado los modelos de vectores de palabras, bien sólo con Wikipedia (W) o como combinación de ésta con los tweets del corpus construido (W+T). La segunda columna indica cómo se ha entrenado el clasificador de polaridad a partir de los textos etiquetados vectorizados con los modelos generados en el paso previo, bien sólo usando los datos de entrenamiento proporcionados por la organizacion (TASS) o incorporando los etiquetados a partir de emoticonos (TASS+T).

Como podemos observar, el uso de una colección de tweets para ampliar la capacidad de representar un modelo basado en vectores de palabras mejora sensiblemente al generado solamente con Wikipedia, pasando de 61,31 % de ajuste a un 62,39 %. En cambio, utilizar los tweets capturados para la fase de entrenamiento supervisado no lleva sino a una caída del rendimiento del sistema.

Tabla 1: Resultados obtenidos sobre el conjunto *full*

w2v	**SVM**	**Accuracy**	**Macro**-F1
W	TASS	61,31 %	48,55 %
W+T	TASS	62,39 %	50,44 %
W	TASS+T	49,28 %	40,20 %
W+T	TASS+T	53,72 %	44,10 %

Esto nos lleva a plantearnos la pregunta de qué ocurriría si utilizáramos sólo los tweets recopilados para generar un modelo de vectores de palabras. Los resultados que se obtienen son un 59,05 % de ajuste y un 44,43 % de F1. No cabe duda de que conviene explorar el uso de modelos de generación de características a partir de vectores de palabras.

Estos resultados mejoran nuestros datos del año pasado, en los que obtuvimos un ajuste del 61,19 % combinando vectores de palabras (Word2Vec) y vectores de documentos (Doc2Vec).

6 Conclusiones y trabajo futuro

A partir de los resultados obtenidos, encontramos que resulta interesante la incorporación de texto no formal (tweets) para la generación de los modelos de palabras, lo cual tiene su sentido en una tarea de clasificación que, precisamente, trabaja sobre textos no formales que tienen la misma red social como fuente. En cambio, el considerar que los emoticonos en un tweet pueden ayudar a un clasificador como SVM a mejorar en la determinación de la polaridad ha resultado una hipótesis fallida. Esto puede entenderse echando un vistazo a algunos de los tweets capturados por el sistema, donde se evidencia la dificultad, incluso para una persona, de poner en contexto el sentido del tweet y su consideración como positivo o negativo si no disponemos de un emoticono asociado.

Como trabajo futuro nos proponemos diseñar una red neuronal profunda más elaborada, pero que parta también de textos de entrenamiento tanto formales como no formales, si bien teniendo en cuanta información lingüística más avanzada como la sintáctica, en lugar de trabajar con simples bolsas de palabras. También queremos explorar el uso

de redes de este tipo en el proceso de clasficación en sí, y no sólo en la generación de características. Una posibilidad es utilizar una red de tipo DBN (Deep Belief Network) (Hinton y Salakhutdinov, 2006) en la que se añade una última fase donde se realiza el etiquetado de los ejemplos.

Bibliografía

Bengio, Yoshua. 2009. Learning deep architectures for ai. *Foundations and trends in Machine Learning*, 2(1):1–127.

Brooke, Julian, Milan Tofiloski, y Maite Taboada. 2009. Cross-linguistic sentiment analysis: From english to spanish. En Galia Angelova Kalina Bontcheva Ruslan Mitkov Nicolas Nicolov, y Nikolai Nikolov, editores, *RANLP*, páginas 50–54. RANLP 2009 Organising Committee / ACL.

Collobert, Ronan y Jason Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. En *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, páginas 160–167, New York, NY, USA. ACM.

Díaz-Galiano, M.C. y A. Montejo-Ráez. 2015. Participación de SINAI DW2Vec en TASS 2015. En *In Proc. of TASS 2015: Workshop on Sentiment Analysis at SEPLN. CEUR-WS.org*, volumen 1397.

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 *In Proc. of the TASS workshop at SEPLN 2013*.

García-Cumbreras, Miguel Ángel, Julio Villena-Román, Eugenio Martínez-Cámara, Manuel Carlos Díaz-Galiano, Mª. Teresa Martín-Valdivia, y L. Alfonso Ureña-López. 2016. Overview of tass 2016. En *Proceedings of TASS 2016: Workshop on Sentiment Analysis at SEPLN co-located with the 32nd SEPLN Conference (SEPLN 2016)*, Salamanca, Spain, September.

Hinton, Geoffrey E y Ruslan R Salakhutdinov. 2006. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507.

Hurtado, Lluís F y Ferran Pla. 2014. Elirf-upv en tass 2014: Análisis de sentimientos, detección de tópicos y análisis de sentimientos de aspectos en twitter. En *In Proc. of the TASS workshop at SEPLN 2014*.

Hurtado, Lluís-F, Ferran Pla, y Davide Buscaldi. 2015. Elirf-upv en tass 2015: Análisis de sentimientos en twitter. En *In Proc. of TASS 2015: Workshop on Sentiment Analysis at SEPLN. CEUR-WS.org*, volumen 1397, páginas 35–40.

Mikolov, Tomas, Kai Chen, Greg Corrado, y Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Montejo-Ráez, A., M.A. García-Cumbreras, y M.C. Díaz-Galiano. 2014. Participación de SINAI Word2Vec en TASS 2014. En *In Proc. of the TASS workshop at SEPLN 2014*.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, y others. 2011. Scikit-learn: Machine learning in python. *The Journal of Machine Learning Research*, 12:2825–2830.

Saralegi Urizar, Xabier y Iñaki San Vicente Roncal. 2012. Tass: Detecting sentiments in spanish tweets. En *TASS 2012 Working Notes*.

Socher, Richard, Jeffrey Pennington, Eric H. Huang, Andrew Y. Ng, y Christopher D. Manning. 2011. Semi-supervised recursive autoencoders for predicting sentiment distributions. En *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, páginas 151–161, Stroudsburg, PA, USA. Association for Computational Linguistics.

TASS 2016: Workshop on Sentiment Analysis at SEPLN, septiembre 2016, pág. 47-51

ELiRF-UPV en TASS 2016: Análisis de Sentimientos en Twitter

ELiRF-UPV at TASS 2016: Sentiment Analysis in Twitter

Lluís-F. Hurtado y Ferran Pla
Universitat Politècnica de València
Camí de Vera s/n
46022 València
{lhurtado, fpla}@dsic.upv.es

Resumen: En este trabajo se describe la participación del equipo del grupo de investigación ELiRF de la Universitat Politècnica de València en el Taller TASS2016. Este taller es un evento enmarcado dentro de la XXXII edición del Congreso Anual de la Sociedad Española para el Procesamiento del Lenguaje Natural. Este trabajo presenta las aproximaciones utilizadas para las dos tareas planteadas en el taller, los resultados obtenidos y una discusión de los mismos. Nuestra participación se ha centrado principalmente en explorar diferentes aproximaciones para combinar un conjunto de sistemas con lo que se ha obtenido los mejores resultados en ambas tareas.
Palabras clave: Twitter, Análisis de Sentimientos.

Abstract: This paper describes the participation of the ELiRF research group of the Universitat Politècnica de València at TASS2016 Workshop. This workshop is a satellite event of the XXXII edition of the Annual Conference of the Spanish Society for Natural Language Processing. This work describes the approaches used for the two tasks of the workshop, the results obtained and a discussion of these results. Our participation has focused primarily on exploring different approaches for combining a set of systems. Using these approaches we have achieved the best results in both tasks.
Keywords: Twitter, Sentiment Analysis.

1. Introducción

El Taller de Análisis de Sentimientos (TASS) en sus cinco ediciones ha venido planteando tareas relacionadas con el análisis de sentimientos en Twitter. El objetivo principal es el de comparar y evaluar diferentes aproximaciones a estas tareas. Además, desarrolla recursos de libre acceso, básicamente, corpora anotados con polaridad, temática, tendencia política, aspectos, que son de gran utilidad para la comparación de diferentes aproximaciones a las tareas propuestas.

En esta quinta edición del TASS se proponen dos tareas de ediciones anteriores (García-Cumbreras et al., 2016): 1) Determinación de la polaridad en tweets, con diferentes grados de intensidad en la polaridad: 6 etiquetas y 4 etiquetas y 2) Determinación de la polaridad de los aspectos en el corpus STOMPOL. Este corpus consta de un conjunto de tweets sobre diferentes aspectos pertenecientes al dominio de la política.

El presente artículo resume la participación del equipo ELiRF-UPV de la Universitat Politècnica de València en todas las tareas planteadas en este taller. Primero se describen las aproximaciones y recursos utilizados en cada tarea. A continuación se presenta la evaluación experimental realizada y los resultados obtenidos. Finalmente se muestran las conclusiones y posibles trabajos futuros.

2. Descripción de los sistemas

Los sistemas presentados en el TASS 2016 se basan en el sistema desarrollado en la edicion anterior del TASS 2015 (Hurtado, Pla, y Buscaldi, 2015). Muchas de las características y recursos de este sistema fueron utilizados en las ediciones en las que nuestro equipo ha participado (Pla y Hurtado, 2013) (Hurtado y Pla, 2014) . El preproceso de los

ISSN 1613-0073

tweets utiliza la estrategia descrita en el trabajo del TASS 2013 (Pla y Hurtado, 2013). Esta consiste básicamente en la adaptación para el castellano del tokenizador de tweets *Tweetmotif* (Connor, Krieger, y Ahn, 2010). También se ha usado *Freeling* (Padró y Stanilovsky, 2012)[1] como lematizador, detector de entidades nombradas y etiquetador morfosintáctico, con las correspondientes modificaciones para el dominio de Twitter. Usando esta aproximación, la tokenización ha consistido en agrupar todas las fechas, los signos de puntuación, los números y las direcciones web. Se han conservado los hashtags y las menciones de usuario. Se ha considerado y evaluado el uso de palabras y lemas como tokens así como la detección de entidades nombradas.

Todas las tareas se han abordado como un problema de clasificación. Se han utilizado Máquinas de Soporte Vectorial (SVM) por su capacidad para manejar con éxito grandes cantidades de características. En concreto usamos dos librerías (*LibSVM*[2] y *LibLinear*[3]) que han demostrado ser eficientes implementaciones de SVM que igualan el estado del arte. El software está desarrollado en *Python* y para acceder a las librerías de SVM se ha utilizado el toolkit *scikit-learn*[4]. (Pedregosa et al., 2011).

En este trabajo se ha explotado la técnica de combinación de diferentes configuraciones de clasificadores para aprovechar su complementariedad. Se ha utilizado la técnica de votación simple utilizada en trabajos anteriores (Pla y Hurtado, 2013) (Pla y Hurtado, 2014b) pero en este caso extendiéndola a un número mayor de clasificadores, con diferentes parámetros y características (palabras, lemas, n-gramas de palabras y lemas) así como estrategias de combinación alternativas.

Cada tweet se ha representado como un vector que contiene los coeficientes tf-idf de las características consideradas. En toda la experimentación realizada, las características y los parámetros de los clasificadores se han elegido mediante una validación cruzada de 10 iteraciones (10-fold cross-validation) sobre el conjunto de entrenamiento.

[1]http://nlp.lsi.upc.edu/freeling/
[2]http://www.csie.ntu.edu.tw/~cjlin/libsvm/
[3]http://www.csie.ntu.edu.tw/~cjlin/liblinear/
[4]http://scikit-learn.org/stable/

3. *Tarea 1: Análisis de sentimientos en tweets*

Esta tarea consiste en determinar la polaridad de los tweets y la organización ha definido dos subtareas. La primera distingue seis etiquetas de polaridad: N y N+ que expresan polaridad negativa con diferente intensidad, P y P+ para la polaridad positiva con diferente intensidad, NEU para la polaridad neutra y NONE para expresar ausencia de polaridad. La segunda sólo distinguen 4 etiquetas de polaridad: N, P, NEU y NONE.

El corpus proporcionado por la organización del TASS consta de un conjunto de entrenamiento, compuesto por 7219 tweets etiquetados con la polaridad usando seis etiquetas, y un conjunto de test, de 60798 tweets, al cual se le debe asignar la polaridad. La distribución de tweets según su polaridad en el conjunto de entrenamiento se muestra en la Tabla 1.

Polaridad	# tweets	%
N	1335	18.49
N+	847	11.73
NEU	670	9.28
NONE	1483	20.54
P	1232	17.07
P+	1652	22.88
TOTAL	7219	100

Tabla 1: Distribución de tweets en el conjunto de entrenamiento según su polaridad.

A partir de la tokenización propuesta se realizó un proceso de validación cruzada (10-fold cross validation) para determinar el mejor conjunto de características y los parámetros del modelo. Como características se probaron diferentes tamaños de n-gramas de palabras y de lemas. También se exploró la combinación de los modelos mediante diferentes técnicas de votación para aprovechar su complementariedad y mejorar las prestaciones finales. Algunas de éstas técnicas proporcionaron mejoras significativas sobre el mismo conjunto de datos, como se muestra en (Pla y Hurtado, 2014b). En todos los casos se han utilizado diccionarios de polaridad, tanto de lemas (Saralegi y San Vicente, 2013), como de palabras (Martínez-Cámara et al., 2013) y el diccionario *Afinn* (Hansen et al., 2011) traducido automáticamente del inglés al castellano.

Se han considerado dos alternativas para abordar la tarea:

- **run1** La primera alternativa combina mediante un sistema de votación ponderada la salida de 192 clasificadores basados en el uso de SVM. La diferencia entre los clasificadores radica en el preprocesado y la tokenización utilizada, las características seleccionadas y los valores de los parámetros del propio modelo SVM.

  En concreto se realizaron todas las combinaciones posibles entre 8 tokenizaciones (lemas o palabras, detectar NE o no, detectar menciones a usuarios y hashtags, ...); 4 conjuntos distinto de características (palabras o bigramas con y sin diccionarios de polaridad) y 6 valores distintos del parámetro c del modelo SVM con kernel lineal.

  La clase asignada a cada tweet t viene determinada por la siguiente fórmula.

$$\hat{c} = \underset{c \in \mathcal{C}}{\operatorname{argmax}}(N_t(c) \cdot P(c)) \qquad (1)$$

  Donde $\mathcal{C}$ es el conjunto de todas las clases, $N_t(c)$ es el número de clasificadores que asignan la clase c al tweet t, y $P(c)$ es la probabilidad a priori de la clase c calculada utilizando el corpus de entrenamiento.

- **run2** La segunda alternativa explora la combinación de modelos mediante el aprendizaje de un metaclasificador. Utilizando las salidas de los mismos 192 clasificadores que en el run anterior, se ha aprendido un segundo modelo SVM que sirve para proporcionar la nueva salida combinada. Se ha destinado una parte del corpus de entrenamiento para ajustar los parámetros del metamodelo. Esta aproximación es la misma que la utilizada en la edición del TASS 2015.

Para la subtarea de 4 etiquetas el **run1** se ha aprendido utilizando el corpus de aprendizaje con 4 etiquetas mientras que el **run2**, dada la complejidad del ajuste de parámetros del metamodelo se ha optado por adaptar el resultado de la subtarea de 6 etiquetas uniendo P y P+ como P y N y N+ como N.

En la Tabla 2 se muestran los valores de Accuracy obtenidos para las dos subtareas. Los sistemas presentados han obtenido las dos primeras posiciones en las dos subtareas consideradas.

	Run	Accuracy
6-ETIQUETAS	run1	0.662
	run2	0.673
4-ETIQUETAS	run1	0.707
	run2	0.721

Tabla 2: Resultados oficiales del equipo *ELiRF-UPV* en la Tarea 1 de la competición TASS-2016 sobre el conjunto de test para 6 y 4 etiquetas.

4. Tarea 2: Análisis de Polaridad de Aspectos en Twitter

Esta tarea consiste en asignar la polaridad a los aspectos que aparecen marcados en el corpus. Una de las dificultades de la tarea consiste en definir qué contexto se le asigna a cada aspecto para poder establecer su polaridad. Para un problema similar, detección de la polaridad a nivel de entidad, en la edición del TASS 2013, propusimos una segmentación de los tweets basada en un conjunto de heurísticas (Pla y Hurtado, 2013). Esta aproximación también se utilizó para la tarea de detección de la tendencia política de los usuarios de Twitter (Pla y Hurtado, 2014a) y para este caso proporcionó buenos resultados. En este trabajo se propone una aproximación más simple que consiste en determinar el contexto de cada aspecto a través de una ventana fija definida a la izquierda y derecha de la instancia del aspecto. Esta aproximación es la que se utilizó en nuestro sistema del TASS 2015 la cual utiliza ventanas de diferente longitud. La longitud de la ventana óptima se ha determinado experimentalmente sobre el conjunto de entrenamiento mediante una validación cruzada. Para entrenar nuestro sistema, se ha considerado el conjunto de entrenamiento únicamente, se han determinado los segmentos para cada aspecto y se ha seguido una aproximación similar a la Tarea 1.

El corpus de la tarea, corpus *STOMPOL*, se compone de un conjunto de tweets relacionados con una serie de aspectos políticos (como economía, sanidad, etc.) enmarcados en la campaña política de las elecciones andaluzas de 2015. Cada aspecto se relaciona con una o varias entidades que se corresponden

con uno de los principales partidos políticos en España (PP, PSOE, IU, UPyD, Cs y Podemos). El corpus consta de 1.284 tweets, y ha sido dividido en un conjunto de entrenamiento (784 tweets) y un conjunto de evaluación (500 tweets).

4.1. Aproximación y resultados

A continuación presentamos una pequeña descripción de las características de nuestro sistema así como el proceso seguido en la fase de entrenamiento. El sistema utiliza un clasificador basado en SVM. Para aprender los modelos sólo se utiliza el conjunto de entrenamiento proporcionado para la tarea y los diccionarios de polaridad previamente descritos. Antes de abordar el entrenamiento se determinan los segmentos de tweet que constituyen el contexto de cada una de los aspectos presentes. Se ha tenido en cuenta tres tamaños de ventana de longitudes 5, 7 y 10 palabras a la izquierda y derecha del aspecto. Cada uno de los segmentos se tokeniza y se utiliza Freeling para determinar sus lemas y ciertas entidades. A continuación se aprenden diferentes modelos combinando tamaños de ventana, parámetros del modelo y diferentes características (palabras, lemas, NE, etc). Mediante validación cruzada se elige el mejor modelo. Para esta tarea sólo hemos presentado un modelo.

	Run	Accuracy
STOMPOL	run1	0.633

Tabla 3: Resultados oficiales del equipo *ELiRF-UPV* en la Tarea 2 de la competición TASS-2016 para el corpus STOMPOL.

En la Tabla 3 se presentan los resultados obtenidos para la Tarea 2 con lo que nuestra aproximación ha obtenido la primera posición en dicha tarea.

5. Conclusiones y trabajos futuros

En este trabajo se ha presentado la participación del grupo ELiRF-UPV en las 2 tareas planteadas en TASS 2016. Nuestro equipo ha utilizado aproximaciones basadas en máquinas de soporte vectorial y se ha centrado principalmente en combinar diferentes sistemas.

Haciendo un análisis del número de participantes y de los resultados obtenidos en las dos últimas ediciones del TASS, creemos que se está cerca de alcanzar los mejores resultados posibles en la tarea de Análisis de sentimientos tal y como se ha venido planteando hasta el momento.

A la vista de los buenos resultados que se han obtenido mediante la combinación de sistemas, como trabajo futuro nos planteamos desarrollar nuevos métodos de combinación de sistemas más sofisticados así como la inclusión de otros paradigmas de clasificación más hetereogéneos (distintos de los SVM) para aumentar la complementariedad de los sistemas combinados.

Además, se pretende extender el sistema para otros idiomas. El sistema descrito ya ha sido utilizado, con ligeras modificaciones, en tareas de análisis de sentimientos para el Inglés en la competición Semeval (Martínez, Pla, y Hurtado, 2016) aunque con resultados no tan satisfactorios como en las tareas del TASS.

Agradecimientos

Este trabajo ha sido parcialmente subvencionado por el MINECO mediante el proyecto ASLP-MULAN: Audio, Speech and Language Processing for Multimedia Analytics (TIN2014-54288-C4-3-R).

Bibliografía

Connor, Brendan O, Michel Krieger, y David Ahn. 2010. Tweetmotif: Exploratory search and topic summarization for twitter. En William W. Cohen y Samuel Gosling, editores, *Proceedings of the Fourth International Conference on Weblogs and Social Media, ICWSM 2010, Washington, DC, USA, May 23-26, 2010*. The AAAI Press.

García-Cumbreras, Miguel Ángel, Julio Villena-Román, Eugenio Martínez-Cámara, Manuel Carlos Díaz-Galiano, Mª. Teresa Martín-Valdivia, y L. Alfonso Ureña-López. 2016. Overview of tass 2016. En *Proceedings of TASS 2016: Workshop on Sentiment Analysis at SEPLN co-located with the 32nd SEPLN Conference (SEPLN 2016)*, Salamanca, Spain, September.

Hansen, Lars Kai, Adam Arvidsson, Finn Årup Nielsen, Elanor Colleoni, y Michael Etter. 2011. Good friends, bad news-affect and virality in twitter. En

Future information technology. Springer, páginas 34–43.

Hurtado, Lluís F., Ferran Pla, y Davide Buscaldi. 2015. Elirf-upv en tass 2015: Análisis de sentimientos en twitter. En *SEPLN*.

Hurtado, LLuís_F y Ferran Pla. 2014. Elirf-upv en tass 2014: Análisis de sentimientos, detección de tópicos y análisis de sentimientos de aspectos en twitter. En *TASS2014*.

Martínez, Víctor, Ferran Pla, y Lluís-F Hurtado. 2016. Dsic-elirf at semeval-2016 task 4: Message polarity classification in twitter using a support vector machine approach.

Martínez-Cámara, E., M. T. Martín-Valdivia, M. D. Molina-gonzález, y L. A. Ureña-lópez. 2013. Bilingual Experiments on an Opinion Comparable Corpus. En *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, página 87–93.

Padró, Lluís y Evgeny Stanilovsky. 2012. Freeling 3.0: Towards wider multilinguality. En *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)*, Istanbul, Turkey, May. ELRA.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, y E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Pla, Ferran y Lluís-F Hurtado. 2013. Tass-2013: Análisis de sentimientos en twitter. En *Proceedings of the TASS workshop at SEPLN 2013*. IV Congreso Español de Informática.

Pla, Ferran y Lluís-F. Hurtado. 2014a. Political tendency identification in twitter using sentiment analysis techniques. En *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, páginas 183–192, Dublin, Ireland, August. Dublin City University and Association for Computational Linguistics.

Pla, Ferran y Lluís-F. Hurtado. 2014b. Sentiment analysis in twitter for spanish. En Elisabeth Métais Mathieu Roche, y Maguelonne Teisseire, editores, *Natural Language Processing and Information Systems*, volumen 8455 de *Lecture Notes in Computer Science*. Springer International Publishing, páginas 208–213.

Saralegi, Xabier y Iñaki San Vicente. 2013. Elhuyar at tass 2013. En *Proceedings of the TASS workshop at SEPLN 2013*. IV Congreso Español de Informática.

TASS 2016: Workshop on Sentiment Analysis at SEPLN, septiembre 2016, pág. 53-57

GTI at TASS 2016: Supervised Approach for Aspect Based Sentiment Analysis in Twitter*

GTI en TASS 2016: Una aproximación supervisada para el análisis de sentimiento basado en aspectos en Twitter

Tamara Álvarez-López, Milagros Fernández-Gavilanes, Silvia García-Méndez, Jonathan Juncal-Martínez, Francisco Javier González-Castaño
GTI Research Group, AtlantTIC
University of Vigo, 36310 Vigo, Spain
{talvarez,mfgavilanes,sgarcia,jonijm}@gti.uvigo.es, javier@det.uvigo.es

Resumen: Este artículo describe la participación del grupo de investigación GTI, del centro AtlantTIC, perteneciente a la Universidad de Vigo, en el TASS 2016. Este taller es un evento enmarcado dentro de la XXXII edición del Congreso Anual de la Sociedad Española para el Procesamiento del Lenguaje Natural. En este trabajo se propone una aproximación supervisada, basada en clasificadores, para la tarea de análisis de sentimiento basado en aspectos. Mediante esta técnica hemos conseguido mejorar las prestaciones de ediciones anteriores, obteniendo una solución acorde con el estado del arte actual.
Palabras clave: Análisis de sentimiento, aspectos, SVM, aprendizaje automático, Twitter

Abstract: This paper describes the participation of the GTI research group of AtlantTIC, University of Vigo, in TASS 2016. This workshop is framed within the XXXII edition of the Annual Congress of the Spanish Society for Natural Language Processing event. In this work we propose a supervised approach based on classifiers, for the aspect based sentiment analysis task. Using this technique we managed to improve the performance of previous years, obtaining a solution reflecting the actual state-of-the-art.
Keywords: Sentiment analysis, aspects, SVM, machine learning, Twitter

1 Introduction

The social media activity is being profused in the recent years, users post opinions and comments in Twitter and in other social platforms. Due to this, there is a huge amount of information available that could be useful for business, in order to design marketing campaigns or to apply any kind of business analysis.

As a consequence, the research on text mining and also on the field of *Sentiment Analysis* (SA) has grown considerably these days. SA is the part of *Natural Language Processing* (NLP) responsible for determining the polarity of a text or a whole sentence. The SA applied to Twitter has to be conducted in a restricted scenario due to the maximum length of the post. However, tweets have other elements we have to consider, like hashtags, mentions and retweets. More concretely, *aspect-based sentiment analysis* (ABSA) consists of extracting opinions, i.e. determining the sentiment polarity, from specific entities in the text (Liu, 2012). Therefore, this task becomes a challenge on the field of NLP.

The TASS Workshop (García-Cumbreras et al., 2016) and the SEPLN conference offer an opportunity for participants to know about the latest advances on the field of NLP for Spanish language.

Many approaches applied to SA can be found in the literature, where it is possible to distinguish between knowledge based approaches (Brooke, Tofiloski, and Taboada, 2009; Fernández-Gavilanes et al., 2016), using grammars and thesaurus and others based on machine learning approaches (Mo-

* This work was partially supported by the Ministerio de Economía y Competitividad under project COINS (TEC2013-47016-C2-1-R) and by Xunta de Galicia (GRC2014/046).

hammad, Kiritchenko, and Zhu, 2013). In the last years we can also find deep learning approaches (Bengio, 2009), applied to this task.

We present our supervised *machine learning* (ML) system which consists of a *Support Vector Machine* (SVM) classifier. Our objective is to conduct the SA process at an aspect level, task 2, determining the polarity of a specific given part of a sentence.

The article is structured as follows. Section 2 is a review of the research involving SA in the Twitter domain. Then, the Section 3 describes the applied approach and the implemented system. In Section 4, we show the experimental results of our system. Finally, in Section 5 we present the conclusions and future works.

2 Related work

A large amount of literature related to *Opinion Mining* (OM) and SA can be found (Pang and Lee, 2008; Martínez-Cámara et al., 2016). Most of the systems are applied to Twitter. However others are applied to social media platforms within the micro-blog context. Due to this, the approaches are varied technically and in connection with the purpose.

Two main approaches exist in SA: supervised and unsupervised learning ones. Supervised systems implement classification methods like SVM, *Logistic Regression* (LR), *Conditional Random Fields* (CRF), *K-Nearest Neighbors* (KNN), etc. Cui, Mittal, and Datar (2006) affirmed that SVM are more appropriate for sentiment classification than generative models, due to their capability for working with ambiguity, that is, dealing with mixed feelings. Supervised algorithms are used when the number of classes, as well as the representative members of each class, are known.

Unsupervised systems are based on linguistic knowledge like lexicons, and syntactic features in order to infer the polarity (Paltoglou and Thelwall, 2012). These last techniques represent a more effective approach in the cross-domain context and for multilingual applications. The unsupervised classification algorithms do not work with a training set, in contrast, some of them use clustering algorithms in order to distinguish groups (Li and Liu, 2010).

As noted earlier, the special case of applying SA to Twitter has been fully addressed (Pak and Paroubek, 2010; Han and Baldwin, 2011). Within the chosen solutions, we highlight the text normalization approach (Fabo, Cuadros, and Etchegoyhen, 2013) and the use of key elements in classification approach (Wang et al., 2011). Others hold the advantages of using deep learning techniques in this task (dos Santos and Gatti, 2014).

According to the purpose of the developed systems, it is possible to find applications like classification of product reviews and political sentiment and election results prediction (Bermingham and Smeaton, 2011), among others.

3 System Overview

In this section we make a brief description of the system submitted for Task 2: Aspect-based sentiment analysis. We developed a supervised system, based on a SVM classifier using different features. In the next subsections we explain the different steps required.

3.1 Preprocessing

Before applying any supervised approach to our corpus, some preprocessing is needed. First of all, we have to normalize the text, since in Twitter language we can find abbreviations, mentions, hashtags, URLs or misspellings. In order to do that, we replace the URLs with the "URL" tag and we replace the abbreviations or misspellings with the correct entire word. For mentions and hashtags, we keep them unchanged but deleting the "@" or "#" symbols. Moreover, when a hashtag is composed of several words, we split and treat them as different tokens.

After this, a lexical analysis is carried out. It consists of lemmatization and POS tagging, which are performed by means of Freeling tool (Atserias et al., 2006).

Once we have analysed lexically the texts, we decided to separate the sentences by the different aspects. For doing that, the scope of each aspect is determined, applying the following rules, which are adapted from our English aspect based sentiment anaylisis system (Alvarez-López et al., 2016)

- If there is only one aspect in the sentence, we keep the sentence unchanged, and introduce it entirely as input for the next step.

- If there are multiple aspects, we separate the sentences by punctuation marks, conjunctions or other aspects found.
- If there are several aspects with no words between them, we consider that they belong to the same context, and assign the same polarity to all of them.

3.2 SVM classifier

In this section we describe the strategy followed to determine the sentiment (positive, negative or neutral) for each aspect predefined in corpus.

We develop a SVM classifier, using the *libsvm* library (Chang and Lin, 2011). The inputs for the SVM will be the sentences separated by contexts, as explained in the previous subsection. The features extracted are the following:

- *Word tokens* of nouns, adjectives and verbs in the sentence.
- *Lemmas* of verbs, nouns and adjectives that appear in each sentence.
- *POS tags* of nouns, adjectives and verbs.
- *N-grams* of different length, grouping the words in each sentence.
- *Aspects* appearing in the sentence. We join "aspect"-"entity", defined in each target as a feature.
- *Negations.* We create a negation dictionary, which contains several particles indicating negation, such as "*no*", "*nunca*", etc.

The previous features are all binary ones, assigning the value 1 if the current feature is present in the tweet and the value 0, if not.

4 Experimental Results

The Task 2: Sentiment Analysis at the aspect level consists of assigning a polarity label to each aspect, which were initially marked in the STOMPOL corpus (Martínez-Cámara et al., 2016) raised by the TASS organization. In this way, this corpus provides both polarity labels and the identification of the aspects that appear in each tweet. The aim is to be able to correctly assign to each aspect a positive, negative or neutral polarity.

In this regard, the STOMPOL corpus consists of a set of Spanish tweets related to a number of political issues, such as health or economy, among others. These issues are framed in the political campaign of Andalusian elections in 2015, where each aspect relates to one or several entities that correspond to one of the main political parties in Spain (PP, PSOE, IU, UPyD, Cs and Podemos). The corpus is composed by 1,284 tweets, and has been divided into a training set (784 tweets) and a set of evaluation (500 tweets).

In order to evaluate the performance of the various features for polarity classification at an aspect-based level, we perform a series of ablation experiments as shown in Table 1. We start with the word token baseline classifier, and then add all four sets of features that help to increase performance as measured by accuracy. As we might expect, including the aspect feature has the most marked effect on the performance of polarity classification, although all the features contributed to improving overall performance on STOMPOL corpus.

Type	Accuracy	Improvement
Word token	56.12	
+Lemmas	57.64	+1.52%
+POS tags	58.26	+0.62%
+Aspects	59.94	+1.68%
+Negations	60.60	+0.66%

Table 1: Results for polarity feature ablation experiments on STOMPOL corpus

Due to the low participation of research teams in task 2 this year, we decided to compare our proposal to the systems presented this year and also to that ones of last year, because of the use of the same dataset.

For this reason, Table 2 compares results for our approach with different official ones submitted in 2015 and 2016 TASS editions. In this way, we compared our results for a ML approach based on well-known squared-regularised logistic regression with a snippet of length 4 (Lys-2) described in Vilares et al. (2015), a clustering method focused on grouping authors with similar sociolinguistic insights (TID-spark) described in Park (2015), a recurrent neural network composed of a single long short term memory and a logistic function (Lys-1) described in Vilares et al. (2015), a ML approach based on a

SVM with a snipped of length 5,7 and 10 (ELiRF) described in Hurtado, Plà, and Buscaldi (2015), and the best performing run of the actual task 2 TASS edition (ELiRF-UPV).

Experiment	Task edition	Accuracy
ELiRF-UPV	2016	63.3
ELiRF	2015	63.3
GTI	**2016**	**60.6**
LyS-1	2015	59.9
TID-spark	2015	55.7
Lys-2	2015	54.0

Table 2: Results of different approaches in 2015/2016 TASS editions on STOMPOL corpus

Comparing the results, the performance of our current model is close from the top ranking systems of this and last year.

5 Conclusions and future works

This paper describes the participation of the GTI group in the TASS 2016, Task 2: Aspect-Based Sentiment Analysis. We developed a supervised system based on a SVM classifier for the aspect-based sentiment analysis. The performance of our approach has been compared to that ones submitted this year but also to that ones submitted last year. Experimental results suggest that we need to include explore new features, such as word embedding representations or paraphrase (Zhao and Lan, 2015), in order to improve the performance.

As future work we plan to include new features explained before and to develop a new system which combines different ML classification methods. We are also interested in considering different paradigms of heterogeneous classification, such as deep learning to increase the performance.

References

Alvarez-López, T., J. Juncal-Martınez, M. Fernández-Gavilanes, E. Costa-Montenegro, and F. J. González-Castano. 2016. Gti at semeval-2016 task 5: Svm and crf for aspect detection and unsupervised aspect-based sentiment analysis. *Proceedings of SemEval*, pages 306–311.

Atserias, J., B. Casas, E. Comelles, M. González, L. Padró, and M. Padró. 2006. Freeling 1.3: Syntactic and semantic services in an open-source NLP library. In *Proceedings of LREC*, volume 6, pages 48–55.

Bengio, Y. 2009. Learning deep architectures for AI. *Found. Trends Mach. Learn.*, 2(1):1–127, January.

Bermingham, A. and A. F. Smeaton. 2011. On using Twitter to monitor political sentiment and predict election results.

Brooke, J., M. Tofiloski, and M. Taboada. 2009. Cross-linguistic sentiment analysis: From english to spanish. In G. Angelova, K. Bontcheva, R. Mitkov, N. Nicolov, and N. Nikolov, editors, *RANLP*, pages 50–54. RANLP 2009 Organising Committee / ACL.

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.

Cui, H., V. Mittal, and M. Datar. 2006. Comparative experiments on sentiment classification for online product reviews. In *Proceedings of the 21st National Conference on Artificial Intelligence - Volume 2*, AAAI'06, pages 1265–1270. AAAI Press.

dos Santos, C. N. and M. Gatti. 2014. Deep convolutional neural networks for sentiment analysis of short texts. In *COLING*, pages 69–78.

Fabo, P. R., M. Cuadros, and T. Etchegoyhen. 2013. Lexical normalization of spanish tweets with preprocessing rules, domain-specific edit distances, and language models. In *Proceedings of the Tweet Normalization Workshop co-located with 29th Conference of the Spanish Society for Natural Language Processing (SEPLN 2013), Madrid, Spain, September 20th, 2013.*, pages 59–63.

Fernández-Gavilanes, M., T. Álvarez-López, J. Juncal-Martínez, E. Costa-Montenegro, and F. J. González-Castaño. 2016. Unsupervised method for sentiment analysis in online texts. *Expert Systems with Applications*, 58:57–75.

García-Cumbreras, M. A., J. Villena-Román, E. Martínez-Cámara, M. C. Díaz-Galiano, M. T. Martín-Valdivia, and L. A. Ureña-López. 2016. Overview of tass 2016. In *Proceedings of TASS 2016: Workshop on*

Sentiment Analysis at SEPLN co-located with the 32nd SEPLN Conference (SEPLN 2016), Salamanca, Spain, September.

Han, B. and T. Baldwin. 2011. Lexical normalisation of short text messages: Makn sens a #twitter. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 368–378, Stroudsburg, PA, USA. Association for Computational Linguistics.

Hurtado, L. F., F. Plà, and D. Buscaldi. 2015. ELiRF-UPV en TASS 2015: Análisis de sentimientos en Twitter. In *Proceedings of TASS 2015: Workshop on Sentiment Analysis at SEPLN co-located with 31st SEPLN Conference (SEPLN 2015), Alicante, Spain, September 15, 2015.*, pages 75–79.

Li, G. and F. Liu. 2010. A clustering-based approach on sentiment analysis. In *Intelligent Systems and Knowledge Engineering (ISKE), 2010 International Conference on*, pages 331–337. IEEE.

Liu, B. 2012. *Sentiment Analysis and Opinion Mining*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Martínez-Cámara, E., M. A. García-Cumbreras, J. Villena-Román, and J. García-Morera. 2016. Tass 2015 - the evolution of the spanish opinion mining systems. *Procesamiento del Lenguaje Natural*, 56:33–40.

Mohammad, S. M., S. Kiritchenko, and X. Zhu. 2013. Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets. In *Proceedings of the seventh international workshop on Semantic Evaluation Exercises (SemEval-2013)*, Atlanta, Georgia, USA, June.

Pak, A. and P. Paroubek. 2010. Twitter as a corpus for sentiment analysis and opinion mining. In N. C. C. Chair), K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, and D. Tapias, editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may. European Language Resources Association (ELRA).

Paltoglou, G. and M. Thelwall. 2012. Twitter, myspace, digg: Unsupervised sentiment analysis in social media. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 3(4):66.

Pang, B. and L. Lee. 2008. Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.*, 2(1-2):1–135, January.

Park, S. 2015. Sentiment classification using sociolinguistic clusters. In *Proceedings of TASS 2015: Workshop on Sentiment Analysis at SEPLN co-located with 31st SEPLN Conference (SEPLN 2015), Alicante, Spain, September 15, 2015.*, pages 99–104.

Vilares, D., Y. Doval, M. A. Alonso, and C. Gómez-Rodríguez. 2015. Lys at TASS 2015: Deep learning experiments for sentiment analysis on spanish tweets. In *Proceedings of TASS 2015: Workshop on Sentiment Analysis at SEPLN co-located with 31st SEPLN Conference (SEPLN 2015), Alicante, Spain, September 15, 2015.*, pages 47–52.

Wang, X., F. Wei, X. Liu, M. Zhou, and M. Zhang. 2011. Topic sentiment analysis in Twitter: A graph-based hashtag sentiment classification approach. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, pages 1031–1040, New York, NY, USA. ACM.

Zhao, J. and M. Lan. 2015. Ecnu: Leveraging word embeddings to boost performance for paraphrase in Twitter. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 34–39, Denver, Colorado, June. Association for Computational Linguistics.