TASS
Taller de Análisis de Sentimientos en la SEPLN

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	Daedalus S.A.	jvillena@daedalus.es
Janine García-Morera	Daedalus S.A.	jgarcia@daedalus.es
Miguel Á. García Cumbreras	Universidad de Jaén	magc@ujaen.es
Eugenio Martínez Cámara	Universidad de Jaén	emcamara@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 de edición: 2015

Editores:

Julio Villena-Román	Daedalus S.A.	jvillena@daedalus.es
Janine García-Morera	Daedalus S.A.	jgarcia@daedalus.es
Miguel Á. García Cumbreras	Universidad de Jaén	magc@ujaen.es
Eugenio Martínez Cámara	Universidad de Jaén	emcamara@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

Colaboradores

David Vilares Calvo	Universidad de la Coruña (España)
Ferran Pla Santamaría	Universidad de Valencia (España)
Lluís F. Hurtado	Universidad de Valencia (España)
David Tomás	Universidad de Alicante (España)
Yoan Gutierrez Vázquez	Universidad de Alicante (España)
Manuel Montes y Gómez	Instituto Nacional de Astrofísica, Óptica y Electrónica (México)
Luis Villaseñor-Pineda	Instituto Nacional de Astrofísica, Óptica y Electrónica (México)

Comité de programa

Alexandra Balahur	EC-Joint Research Centre (Italia)
José Carlos González-Cristobal	Universidad Politécnica de Madrid (España)
José Carlos Cortizo	Universidad Europea de Madrid (España)
Ana García-Serrano	Universidad Nacional de Educación a Distancia (España)
Jose María Gómez Hidalgo	Optenet (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)
Paloma Martínez Fernández	Universidad Carlos III 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)
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:

- ATTOS (TIN2012-38536-C03-0)
- AROESCU (P11-TIC-7684 MO)
- Ciudad2020 (INNPRONTA IPT-20111006)
- SAM (FP7-611312)

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 2015 han participado 18 equipos, de los que 14 han enviado un artículo describiendo el sistema que han presentado, habiendo sido aceptados los 14 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 2015 tendrá lugar en el seno del XXXI congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural, que se celebrará el próximo mes de septiembre en Alicante (España).

Agosto de 2015
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.

In the 2015 edition of the workshop has participated 18 teams, from which 14 have submitted a description paper of the system submitted. After a revision process, the organizing committee has accepted the 14 papers. The main goal of the revision was the publication of those papers with a minimum of scientific quality.

The 2015 edition will be held at the 31st conference of the Spanish Society for Natural Language Processing, which will take place at Alicante in September.

August 2015
The editors

TASS
Taller de Análisis de Sentimientos en la SEPLN

CEUR Workshop Proceedings

ISSN: 1613-0073

Artículos

Artículos

TASS 2015, septiembre 2015, pp 13-21 recibido 20-07-15 revisado 24-07-15 aceptado 25-07-15

Overview of TASS 2015

Resumen de TASS 2015

Julio Villena Román
Janine García Morera
Daedalus, S.A.
28031 Madrid, Spain
{jvillena, jgarcia, cdepablo}@daedalus.es

Miguel Ángel García Cumbreras
Eugenio Martínez Cámara
M. Teresa Martín Valdivia
L. Alfonso Ureña López
Universidad de Jaén
23071 Jaén, Spain
{magc, emcamara, laurena, maite}@uja.es

Resumen: Este artículo describe la cuarta edición del taller de evaluación experimental TASS 2015, enmarcada dentro del congreso internacional SEPLN 2015. 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 2015, así como el contenido de los corpus utilizados, los participantes en las distintas tareas y los resultados generales obtenidos y el análisis de estos resultados.
Palabras clave: TASS 2015, análisis de opiniones, medios sociales

Abstract: This paper describes TASS 2105, the fourth edition of the Workshop on Sentiment Analysis at SEPLN. The main objective is to promote the research and the development of new algorithms, resources and techniques in the field of sentiment analysis in social media (specifically Twitter), focused on Spanish language. This paper presents the TASS 2015 proposed tasks, the contents of the generated corpora, the participant groups and the results and analysis of them.
Keywords: TASS 2015, 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 of sentiment analysis systems in social media, focused on Spanish language. The fourth edition will be held on September 15th, 2015 at University of Alicante, Spain.

Sentiment analysis (SA) can be defined as the computational treatment of opinion, sentiment and subjectivity in texts (Pang & Lee, 2002). It is a hard task because even humans often disagree on the sentiment of a given text. And it is a harder task when the text has only 140 characters (Twitter messages or tweets).

Text classification techniques, although studied and improved for a longer time, still need more research effort and resources to be able to build better models to improve the current result values. 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 showed 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 2015. 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 2015 experiments are based on three corpus, specifically built for the different editions of the workshop.

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 Spain-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.

Eachtweet 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 both 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).

All tagging has been done semiautomatically: a baseline machine learning model is first run and then all tags are manually checked by human experts. In the case of the polarity at entity level, due to the high volume of data to check, this tagging has just 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 this year 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 sample tweets. The first tweet is only tagged with the global polarity as the text contains no mentions to any entity, but the second one is tagged with both 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 Social-TV corpus

The Social-TV corpus was collected during the 2014 Final of Copa del Rey championship in Spain between Real Madrid and F.C. Barcelona, played on 16 April 2014 at Mestalla Stadium in Valencia. After filtering useless information a subset of 2.773 tweets was selected.

All tweets were manually tagged with the aspects and its sentiment polarity. Tweets may cover more than one aspect.

The list of the 31 aspects that have been defined is the following:

- Afición (supporters)
- Árbitro (referee)
- Autoridades (authorities)
- Entrenador (coach)
- Equipo - Atlético de Madrid (Team- Atlético de Madrid)
- Equipo - Barcelona (Team- Barcelona)
- Equipo - Real Madrid (Team - Real Madrid)
- Equipo (any other team)
- Jugador - Alexis Sánchez (Player - Alexis Sánchez)
- Jugador - Álvaro Arbeloa (Player - Álvaro Arbeloa)
- Jugador - Andrés Iniesta (Player - Andrés Iniesta)
- Jugador - Ángel Di María (Player - Ángel Di Maria)
- Jugador - Asier Ilarramendi (Player - Asier Ilarramendi)
- Jugador - Carles Puyol (Player - Carles Puyol)
- Jugador - Cesc Fábregas (Player - Cesc Fábregas)
- Jugador - Cristiano Ronaldo (Player - Cristiano Ronaldo)
- Jugador - Dani Alves (Player - Dani Alves)
- Jugador - Dani Carvajal (Player - Dani Carvajal)
- Jugador - Fábio Coentrão (Player - Fábio Coentrão)
- Jugador - Gareth Bale (Player - Gareth Bale)
- Jugador - Iker Casillas (Player - Iker Casillas)
- Jugador - Isco (Player - Isco)
- Jugador - Javier Mascherano (Player - Javier Mascherano)
- Jugador - Jesé Rodríguez (Player - Jesé Rodríguez)
- Jugador - José Manuel Pinto (Player - José Manuel Pinto)
- Jugador - Karim Benzema (Player - Karim Benzema)
- Jugador - Lionel Messi (Player - Lionel Messi)
- Jugador - Luka Modric (Player - Luka Modric)
- Jugador - Marc Bartra (Player - Marc Bartra)
- Jugador - Neymar Jr. (Player - Neymar Jr.)
- Jugador - Pedro Rodríguez (Player - Pedro Rodríguez)
- Jugador - Pepe (Player - Pepe)
- Jugador - Sergio Busquets (Player - Sergio Busquets)
- Jugador - Sergio Ramos (Player - Sergio Ramos)

- Jugador - Xabi Alonso (Player - Xabi Alonso)
- Jugador - Xavi Hernández (Player - Xavi Hernández)
- Jugador (any other player)
- Partido (Football match)
- Retransmisión (broadcast)

Sentiment polarity has been tagged from the point of view of the person who writes the tweet, using 3 levels: P, NEU and N. No distinction is made in cases when the author does not express any sentiment or when he/she expresses a no-positive no-negative sentiment.

The Social-TV corpus was randomly divided into training set (1.773 tweets) and test set (1.000 tweets), with a similar distribution of both aspects and sentiments. The training set was released previously and the test corpus was provided without any tagging and has been used to evaluate the results provided by the different systems.

The following figure shows the information of three sample tweets in the training set.

```
<tweet id="456544898791907328">
  <sentiment aspect="Equipo-Real_Madrid" polarity="P">#HalaMadrid
  </sentiment> ganamos sin <sentiment aspect="Jugador-Cristiano_Ronaldo"
   polarity="NEU">Cristiano</sentiment>. .perdéis con <sentiment aspect=
  "Jugador-Lionel_Messi" polarity="N">Messi</sentiment>. Hala <sentiment
   aspect="Equipo-Real_Madrid" polarity="P">Madrid</sentiment>! !!!!!
</tweet>
<tweet id="456544898942906369">
  @nevermind2192 <sentiment aspect="Equipo-Barcelona" polarity="P">Barça
  </sentiment> por siempre!!
</tweet>
<tweet id="456544898951282688">
  <sentiment aspect="Partido" polarity="NEU">#FinalCopa</sentiment>
  Hala <sentiment aspect="Equipo-Real_Madrid" polarity="P">Madrid
  </sentiment>, hala <sentiment aspect="Equipo-Real_Madrid" polarity="P"
  >Madrid</sentiment>, campeón de la <sentiment aspect="Partido"
  polarity="P">copa del rey</sentiment>
</tweet>
```

Figure 2: Sample tweets (Social-TV corpus)

2.3 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 in 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, labor policy...
- Health System (Sanidad): hospitals, public/private health system, drugs, doctors...
- Education (Educacion): state school, private school, scholarships...
- Political party (Propio_partido): anything good (speeches, electoral programme...) or bad (corruption, criticism) related to the entity
- Otros_aspectos (Other aspects): 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 (Cs)
- Unión_Progreso_y_Democracia (UPyD)

Each tweet in the corpus has been manually tagged 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 3 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 3: Sample tweets (STOMPOL corpus)

These three corpora will be made freely available to the community after the workshop. Please send an email to tass@daedalus.es 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.

3 Description of tasks

First of all, we are interested in evaluating the evolution of the different approaches for SA and text classification in Spanish during these years. So, the traditional SA at global level task will be repeated again, reusing the same corpus,

to compare results. Moreover, we want to foster the research in the analysis of fine-grained polarity analysis at aspect level (aspect-based SA, one of the new requirements of the market of natural language processing in these areas). So, two legacy tasks will be repeated again, to compare results, and a new corpus has been created for the second task.

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 (legacy) Task 1: Sentiment Analysis at Global Level

This is the same task as previous editions. 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. Participants have been provided with the training set of the General corpus so that they may train and validate their models. There will be 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, but 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 will be used for the evaluation, to allow for comparison among systems. Accuracy, macroaveraged precision, macroaveraged recall and macroaveraged F1-measure have been used to evaluate each run.

Notice that there are two test sets: the complete set and 1k set, a subset of the first one. The reason is that, to deal with the problem of the imbalanced distribution of labels between the training and test set, a selected test subset containing 1.000 tweets with a similar distribution to the training corpus was extracted to be used for an alternate evaluation of the performance of systems.

3.2 (legacy) Task 2: Aspect-based sentiment analysis

Participants have been provided with a corpus tagged with a series of aspects, and systems must identify the polarity at the aspect-level. Two corpora have been provided: the Social-TV corpus, used in TASS 2014, and the new STOMPOL corpus, collected in 2015 (described above). Both corpora have been splitted into training and test set, the first one for building and validating 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 \t polarity
```

[for the Social-TV corpus]

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

[for the STOMPOL corpus]

Allowed polarity values are P, N and NEU.

For evaluation, a single label combining "aspect-polarity" has been considered. Similarly to the first task, accuracy, macroaveraged precision, macroaveraged recall and macroaveraged F1-measure have been calculated for the global result.

4 Participants and Results

This year 35 groups registered (as compared to 31 groups last year) but unfortunately only 7 groups (14 last year) sent their submissions. The list of active participant groups is shown in Table 2, including the tasks in which they have participated.

Fourteen of the seventeen 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 3.

Group	1	2
LIF	X	
ELiRF	X	X
GSI	X	X
LyS	X	X
DLSI	X	
GTI-GRAD	X	
ITAINNOVA	X	
SINAI-ESMA	X	
CU	X	
TID-spark	X	X
BittenPotato	X	
SINAI_wd2v	X	
DT	X	
GAS-UCR	X	
UCSP	X	
SEDEMO	X	
INGEOTEC	X	
Total groups	17	4

Table 2: Participant groups

Group	Report
ELiRF	ELiRF-UPV en TASS 2015: Análisis de Sentimientos en Twitter
GSI	Aspect based Sentiment Analysis of Spanish Tweets
LyS	LyS at TASS 2015: Deep Learning Experiments for Sentiment Analysis on Spanish Tweets
DLSI	Evaluating a Sentiment Analysis Approach from a Business Point of View
GTI-GRAD	GTI-Gradiant at TASS 2015: A Hybrid Approach for Sentiment Analysis in Twitter
ITAINNOVA	Ensemble algorithm with syntactical tree features to improve the opinion analysis
SINAI-EMMA	SINAI-EMMA: Vectores de Palabras para el Análisis de Opiniones en Twitter
CU	Spanish Twitter Messages Polarized through the Lens of an English System
TID-spark	Sentiment Classification using Sociolinguistic Clusters
BittenPotato	BittenPotato: Tweet sentiment analysis by combining multiple classifiers
SINAI_wd2v	Participación de SINAI DW2Vec en TASS 2015
DT	DeustoTech Internet at TASS 2015: Sentiment analysis and polarity classification in Spanish tweets
UCSP	Comparing Supervised Learning Methods for Classifying Spanish Tweets
INGEOTEC	Sentiment Analysis for Twitter: TASS 2015

Table 3: Participant reports

5 Results

Results for each task are described next.

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 4. Accuracy, macroaveraged precision, macroaveraged recall and macroaveraged F1-measure have been used to evaluate each individual label and ranking the systems.

Run Id	Acc
LIF-Run-3	0.672
LIF-Run-2	0.654
ELiRF-run3	0.659
LIF-Run-1	0.628
ELiRF-run1	0.648
ELiRF-run2	0.658
GSI-RUN-1	0.618
run_out_of_date	0.673
GSI-RUN-2	0.610
GSI-RUN-3	0.608
LyS-run-1	0.552
DLSI-Run1	0.595
Lys-run-2	0.568
GTI-GRAD-Run1	0.592
Ensemble exp1.1	0.535
SINAI-EMMA-1	0.502
INGEOTEC-M1	0.488
Ensemble exp3_emotions	0.549
CU-Run-1	0.495
TID-spark-1	0.462
BP-wvoted-v2_1	0.534
Ensemble exp2_emotions	0.524

BP-voted-v2	0.535
SINAI_wd2v_500	0.474
SINAI_wd2v_300	0.474
BP-wvoted-v1	0.522
BP-voted-v1	0.522
BP-rbf-v2	0.514
Lys-run-3	0.505
BP-rbf-v1	0.494
CU-Run-2-CompMod	0.362
DT-RUN-1	0.560
DT-RUN-3	0.557
DT-RUN-2	0.545
GAS-UCR-1	0.342
UCSP-RUN-1	0.273
BP-wvoted-v2	0.009

Table 4: Results for Task 1, 5 levels, whole test corpus

As previously described, an alternate evaluation of the performance of systems was done using a new selected test subset containing 1.000 tweets with a similar distribution to the training corpus. Results are shown in Table 5.

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 same double evaluation using the whole test corpus and a new selected corpus have been carried out, shown Tables 8 and 9.

Run Id	Acc
ELiRF-run2	0.488
GTI-GRAD-Run1	0.509
LIF-Run-2	0.516
GSI-RUN-1	0.487
GSI-RUN-2	0.48
GSI-RUN-3	0.479
LIF-Run-1	0.481
ELiRF-run1	0.476
SINAI_wd2v	0.389
ELiRF-run3	0.477
INGEOTEC-M1	0.431
Ensemble exp1 1K	0.405
LyS-run-1	0.428
Ensemble exp2 1K	0.384
Lys-run-3	0.430
Lys-run-2	0.434
SINAI-EMMA-1	0.411
CU-Run-1-CompMod	0.419
Ensemble exp3 1K	0.396
TID	0.400
BP-voted-v1	0.408
DLSI-Run1	0.385
CU-Run-2	0.397
BP-wvoted-v1	0.416
BP-rbf-v1	0.418
SEDEMO-E1	0.397
DT-RUN-1	0.407
DT-RUN-2	0.408
DT-RUN-3	0.396
GAS-UCR-1	0.338
INGEOTEC-E1	0.174
INGEOTEC-E2	0.168

Table 5: Results for Task 1, 5 levels, selected 1k corpus

Run Id	Acc
LIF-Run-3	0.726
LIF-Run-2	0.725
ELiRF-run3	0.721
LIF-Run-1	0.710
ELiRF-run1	0.712
ELiRF-run2	0.722
GSI-RUN-1	0.690
run_out_of_date	0.725
GSI-RUN-2	0.679
GSI-RUN-3	0.678
DLSI-Run1	0.655
LyS-run-1	0.664
GTI-GRAD-Run1	0.695
TID-spark-1	0.594
INGEOTEC-M1	0.613
UCSP-RUN-2	0.594
UCSP-RUN-3	0.613
Ensemble exp2_3_SPARK	0.591
UCSP-RUN-1	0.602
CU-RUN-1	0.597
Ensemble exp1_3_SPARK	0.610
UCSP-RUN-1-ME	0.600
BP-wvoted-v1	0.593
BP-voted-v1	0.593
Ensemble exp3_3	0.594
DT-RUN-2	0.625
SINAI_wd2v	0.619
SINAI_wd2v_2	0.613
BP-rbf-v1	0.602
Lys-run-2	0.599
DT-RUN-3	0.608
UCSP-RUN-1-NB	0.560
SINAI_w2v	0.604
UCSP-RUN-1-DT	0.536
CU-Run2-CompMod	0.481
DT-RUN-1	0.490
UCSP-RUN-2-ME	0.479
SINAI_d2v	0.429
GAS-UCR-1	0.446

Table 6: Results for Task 1, 3 levels, whole test corpus

Run Id	Acc

LIF-Run-1	0.632
ELiRF-run2	0.610
LIF-Run-2	0.692
BP-wvoted-v1	0.632
GSI-RUN-1	0.658
GTI-GRAD-Run1	0.674
BP-voted-v1	0.611
LyS-run-1	0.634
TID-spark-1	0.649
DLSI-Run1	0.637
ELiRF-run1	0.645
DT-RUN-1	0.601
GSI-RUN-2	0.646
GSI-RUN-3	0.647
ELiRF-run3	0.595
Ensemble exp3 1K 3	0.614
UCSP-RUN-2	0.586
Ensemble exp2 1K 3	0.611
Ensemble exp1 1K 3	0.503
INGEOTEC-M1	0.595
CU-Run-2-CompMod	0.600
CU-RUN-1	0.578
SINAI_wd2v_2_500	0.641
UCSP-RUN-1	0.582
SINAI_w2v	0.627
UCSP-RUN-3	0.626
SINAI_wd2v	0.633
BP-rbf-v1	0.611
UCSP-RUN-1-NB	0.636
UCSP-RUN-1-ME	0.626
Lys-run-2	0.605
DT-RUN-2	0.583
DT-RUN-3	0.571
UCSP-RUN-1-DR	0.495
UCSP-RUN-2-NB	0.559
UCSP-RUN-2-ME	0.509
DT-RUN-1	0.514
GAS-UCR-1	0.556
SINAI_d2v	0.510

Table 7: Results for Task 1, 3 levels, selected 1k corpus

5.2 Task 2: Aspect-based Sentiment Analysis

Submitted runs and results for Task 2, with the Social-TV and STOMPOL corpus, are shown in Tables 10 and 11. Accuracy, macroaveraged precision, macroaveraged recall and macroaveraged F1-measure have been used to evaluate each individual label and ranking the systems.

Run Id	Acc
GSI-RUN-1	0.635
GSI-RUN-2	0.621
GSI-RUN-3	0.557
ELiRF-run1	0.655
LyS-run-1	0.610
TID-spark-1	0.631
GSI-RUN-1	0.533
Lys-run-2	0.522

Table 10: Results for Task 2, Social-TV corpus

Run Id	Acc
ELiRF-run1	0.633
LyS-run-1	0.599
Lys-run-2	0.540
TID-spark-1	0.557

Table 11: Results for Task 2, STOMPOL corpus

6 Conclusions and Future Work

TASS was the first workshop about SA focused on the processing of texts written in Spanish. Clearly this area receives great attraction from research groups and companies, as this fourth edition has had a greater impact in terms of registered groups, and the number of participants that submitted experiments in 2015 tasks has increased.

Anyway, the developed corpus and gold standards, and the reports from participants will for sure be helpful for other research groups approaching these tasks.

TASS corpora will be released after the workshop for free use by the research community. In 2014 the corpora had been downloaded up to date by more than 60 research groups, 25 out of Spain, by groups coming from academia and also from private companies to use the corpus as part of their product development. We expect to reach a similar impact with this year's corpus.

Acknowledgements

This work has been partially supported by a grant from the Fondo Europeo of Desarrollo Regional (FEDER), ATTOS (TIN2012-38536-C03-0) and Ciudad2020 (INNPRONTA IPT-20111006) projects from the Spanish Government, and AORESCU project (P11-TIC-7684 MO).

References

Villena-Román, Julio; Lana-Serrano, Sara; Martínez-Cámara, Eugenio; González-Cristobal, José Carlos. 2013. *TASS - Workshop on Sentiment Analysis at SEPLN*. Revista de Procesamiento del Lenguaje Natural, 50, pp 37-44. http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/4657.

Villena-Román, Julio; García-Morera, Janine; Lana-Serrano, Sara; González-Cristobal, José Carlos. 2014. *TASS 2013 - A Second Step in Reputation Analysis in Spanish*. Revista de Procesamiento del Lenguaje Natural, 52, pp 37-44. http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/4901.

Vilares, David; Doval, Yerai; Alonso, Miguel A.; Gómez-Rodríguez, Carlos. *LyS at TASS 2014: A Prototype for Extracting and Analysing Aspects from Spanish tweets*. In Proc. of the TASS workshop at SEPLN 2014. 16-19 September 2014, Girona, Spain.

Perea-Ortega, José M. Balahur, Alexandra. *Experiments on feature replacements for polarity classification of Spanish tweets*. In Proc. of the TASS workshop at SEPLN 2014. 16-19 September 2014, Girona, Spain.

Hernández Petlachi, Roberto; Li, Xiaoou. *Análisis de sentimiento sobre textos en Español basado en aproximaciones semánticas con reglas lingüísticas*. In Proc. of the TASS workshop at SEPLN 2014. 16-19 September 2014, Girona, Spain.

Montejo-Ráez, A.; García-Cumbreras, M.A.; Díaz-Galiano, M.C. *Participación de SINAI Word2Vec en TASS 2014*. In Proc. of the TASS workshop at SEPLN 2014. 16-19 September 2014, Girona, Spain.

Hurtado, Lluís F.; Pla, Ferran. *ELiRF-UPV en TASS 2014: Análisis de Sentimientos, Detección de Tópicos y Análisis de Sentimientos de Aspectos en Twitter*. In Proc. of the TASS workshop at SEPLN 2014. 16-19 September 2014, Girona, Spain.

Jiménez Zafra, Salud María; Martínez Cámara, Eugenio; Martín Valdivia, M. Teresa.; Ureña López, L. Alfonso. *SINAI-ESMA: An unsupervised approach for Sentiment Analysis in Twitter*. In Proc. of the TASS workshop at SEPLN 2014. 16-19 September 2014, Girona, Spain.

San Vicente Roncal, Iñaki; Saralegi Urizar, Xabier. *Looking for Features for Supervised Tweet Polarity Classification*. In Proc. of the TASS workshop at SEPLN 2014. 16-19 September 2014, Girona, Spain.

TASS 2015, septiembre 2015, pp 23-28 recibido 20-07-15 revisado 24-07-15 aceptado 29-07-15

DeustoTech Internet at TASS 2015: Sentiment analysis and polarity classification in spanish tweets

DeustoTech Internet en TASS 2015: Análisis de sentimientos y clasificación de polaridad en tweets en español

Juan Sixto Cesteros
DeustoTech–Deusto
Institute of Technology
Universidad de Deusto
48007 Bilbao (Spain)
jsixto@deusto.es

Aitor Almeida
DeustoTech–Deusto
Institute of Technology
Universidad de Deusto
48007 Bilbao (Spain)
aitor.almeida@deusto.es

Diego López de Ipiña
DeustoTech–Deusto
Institute of Technology
Universidad de Deusto
48007 Bilbao (Spain)
dipina@deusto.es

Resumen: Este artículo describe nuestro sistema presentado en el taller de análisis de sentimiento TASS 2015. Nuestro sistema aborda la tarea 1 del workshop, que consiste en realizar un análisis automático de sentimiento para determinar la polaridad global de un conjunto de tweets en español. Para ello, nuestro sistema se basa en un modelo supervisado con máquinas de soporte vectorial lineales en combinación con varios léxicos de polaridad. Se estudia la influencia de las diferentes características linguísticas y de diferentes tamaños de n-gramas en la mejora del algoritmo. Así mismo se presentan los resultados obtenidos, las diferentes pruebas que se han realizado, y una discusión sobre los resultados.
Palabras clave: Análisis de sentimientos, clasificación de la polaridad, Twitter

Abstract: This article describes our system presented at the workshop for sentiment analysis TASS 2015. Our system approaches the task 1 of the workshop, which consists on performing an automatic sentiment analysis to determine the global polarity of a set of tweets in Spanish. To do this, our system is based on a model supervised Linear Support Vector Machines combined with some polarity lexicons. The influence of the different linguistic features and the different sizes of n-grams in improving algorithm performance. Also the results obtained, the various tests that have been conducted, and a discussion of the results are presented.
Keywords: Sentiment Analysis, Polarity Classification, Twitter

1 Introduction

Since the origin of Web 2.0, Internet contains a very large amounts of user-generated information on an unlimited number of topics. Many entities such as corporations or political groups try to learn about that knowledge to know the opinion of users. Social Media platforms such as Facebook or Twitter have proven to be useful for this tasks,due to the very high volume of messages that these platforms generate in real time and the very large number of users that use them everyday.

Faced with this challenge, in the last years the number of the Sentiment Analysis researches has increased appreciably, especially those based in Twitter and microblogging. It should be taken into account that the performance of these researches is language-dependent, reflecting the considerable differences between languages and the difficulty of establish standard linguistic rules (Han, Cook, and Baldwin, 2013).

In this context, the TASS[1] workshop (Villena-Román et al., 2015) is an evaluation workshop for sentiment analysis focused on Spanish language, organized as a satellite event of the annual conference of the Spanish Society for Natural Language Processing (SEPLN)[2]. This paper is focused on the first task of the workshop consist on determining the global polarity of twitter messages.

This paper presents a global polarity clas-

[1] Taller de Análisis de Sentimientos en la SEPLN
[2] http://www.sepln.org/

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

sification in Spanish tweets based on polarity lexicons and linguistic features. It is adapted to Spanish tweet texts, which involve particular linguistic characteristics like short length, limited to 140 characters, slang, spelling and grammatical errors and other user mentions.

The rest of the paper is organized as follows: the sentiment analysis related works are described in Section 2, the developed system's description is presented in Section 3, evaluation and results in Section 4 and conclusion and future work are discussed in Section 5.

2 Related work

There exists a large amount of literature addressing the sentiment analysis field, especially applied to Twitter and microblogging context. General surveys about Opinion Mining and Sentiment Analysis may be found (Pang and Lee, 2008), (Martinez-Camara et al., 2014), although due to the enormous diversity of applications on this field, different approaches to solve problems in numerous scopes have been generated, like user classification (Pennacchiotti and Popescu, 2011), Spam detection in social media (Gao et al., 2010), classification of product reviews (Dave, Lawrence, and Pennock, 2003), demographic studies (Mislove et al., 2011), political sentiment and election results prediction (Bermingham and Smeaton, 2011) and even clinical depression prediction via Twitter (De Choudhury et al., 2013).

Twitter has certain specific characteristics which distinguish them from other social networks, e.g. short texts, @user mentions, #hashtags and retweets. All of these characteristics have been extensively studied (Pak and Paroubek, 2010), (Han and Baldwin, 2011). Some of them have been resolved through the text normalization approach (Ruiz, Cuadros, and Etchegoyhen, 2013) while others have been used as key elements in classification approach (Wang et al., 2011). Indeed, several researches prove that the in-depth knowledge of these characteristics will significantly improve the social media based applications (Jungherr, 2013), (Wang et al., 2013).

For several years we assist to an exponential increase of studies based on sentiment analysis and opinion mining in Twitter. According to the state of art, two main approaches exist in sentiment analysis: supervised learning and unsupervised learning. Supervised systems implement classification models based on classification algorithms, being the most frequent the Support Vector Machine (SVM) (Go, Bhayani, and Huang, 2009), Logistic Regression (LR) (Thelwall, Buckley, and Paltoglou, 2012), Conditional Random Fields (CRF) (Jakob and Gurevych, 2010) and K Nearest Neighbors (KNN) (Davidov, Tsur, and Rappoport, 2010). Unsupervised systems are based on the use of lexicons to calculate the semantic orientation(Turney, 2002) and present a new perspective for classification tasks, most effective in cross-domain and multilingual applications.

During the last TASS workshop in 2014 (Villena-Román et al., 2015), LyS presented a supervised liblinear classifier with several lexicons of Spanish language, whose results are among the best in task 1 (Sentiment Analysis at the tweet level) (Vilares et al., 2014). Further, (San Vicente and Saralegi, 2014) presented a Support Vector Machine (SVM) based on a classifier that merges polarity lexicons with several linguistic features as punctuation marks or negation signs. Finally, the best results in task 1 correspond to (Hurtado and Pla, 2014), who present a Linear-SVM based classifier that addresses the task using a one-vs-all strategy in conjuction with a vectorized list of tf-idf coefficients as text representation.

3 System description

Several tools and datasets have been used during the experiments to develop our final system. Because our system only approaches the Task 1: Sentiment Analysis at global level, this consists in a unique pipeline that reaches the process completely. At the beginning, a naive normalization system is applied to the tweet texts with the purpose to standardize several Twitter own features, like #Hashtags or @User mentions. Then, the Freeling language analysis tool[3] (Padró and Stanilovsky, 2012) is used to tokenize, lemmatize and annotate the texts with part-of-speech tags (pos-tagging).

During this step, based on a list of stop words for Spanish language, this words are annotated to be ignored by polarity ranking steps.

[3] http://nlp.lsi.upc.edu/freeling/

The task has been addressed as an automatic multi-class classification job. For this reason, it has been considered appropriate to focus this problem with a one-vs-all strategy, in a similar way to the presented by (Hurtado and Pla, 2014) in TASS 2014. These binary classifiers have been developed using two different approaches, LinearSVC Machines and Support Vector Regression (SVR) Machines. the comparison of machine-learning based results is shown in Results section.

To represent the text's as vectorized features, two main sources have been used: the polarity lexicon punctuations and the Okapi BM25 ranking function, to represent document's scoring (Robertson et al., 1995). BM25 is a bag-of-words retrieval function that ranks a set of documents based on the query terms appearing in each document. The formula used to implement BM25 in the system is defined below:

$$score(D,Q) = \sum_{i=1}^{n} IDF(q_i) \cdot TF(q_i) \quad (1)$$

$$TF(q_i) = \frac{f(q_i,D) \cdot (k_1+1)}{f(q_i,D) + k_1 \cdot (1 - b + b \cdot \frac{|D|}{avgdl})} \quad (2)$$

$$IDF(q_i) = \log \frac{N - n(q_i) + 0{,}5}{n(q_i) + 0{,}5} \quad (3)$$

To calculate the score of a document D, $f(q_i, D)$ is the frecuency of each word lemma (q_i), $|D|$ is the length of the text D in words and $avgdl$ is the average text length. After several experiments over the training corpus, the free parameters k_1 and b have been optimized to $k_1 = 76$ and $b = 0{,}75$. System develops one BM25 dictionary for each one-vs-all classifier.

In conjunction with the document's score, each tweet has been represented using different polarity lexicons in order to classify them into the six (P+, P, NEU, N, N+ and NONE) and the four (P, N, NEU and NONE) polarities. We use several datasets to score the polarity levels of words and lemmas. Owing to different characteristics of each dataset, such as semantic-orientation values, scores are calculated separately and considered as independent attributes in the system.

- **LYSA Twitter lexicon v0.1**. LYSA is an automatically-built polarity lexicon for Spanish language that was created by downloading messages from Twitter, and includes both negative and positive Spanish words (Vilares et al., 2014). The lexicon entries includes a semantic-orientation values ranged from -5 to 5, making it a good resource for multiple sentiment levels identification.
- **ElhPolar dictionary v1.0**. The ElhPolar polarity lexicon for Spanish was created from different sources, and includes both negative and positive words (Saralegi and San Vicente, 2013).
- **The Spanish Opinion Lexicon (SOL)**. The Spanish Opinion Lexicon (SOL) is composed by 1,396 positive and 3,151 negative words, thus in total SOL has 4,547 opinion words[4] (Martínez-Cámara et al., 2013). The lexicon has been elaborated from the Bing Liu's word list using Reverso as translator (M. and L., 2004).
- **Negation Words List**. A list of negation spanish words has been created during the experiments. This list is used as a text feature in order to detect negative sentences and possible polarity inversions.

We also consider other text characteristics as classifier features, like text length in words quantity or a list of sentiments represented by emoticons using the Wikipedia's list of emoticons[5]. To conclude the system's prediction, another automatic classifier has been implemented, trained with the predictions of the binary results to select one label.

4 Results

Our results are relative to the Task 1: Sentiment Analysis at global level of TASS 2015. This task consists on performing an automatic sentiment analysis to determine the global polarity of each message in the provided corpus. 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). Also there are two test sets: complete set and 1k

[4] http://sinai.ujaen.es/sol/

[5] https://en.wikipedia.org/wiki/List_of_emoticons

set, a subset of the first one containing only 1000 tweets with a similar distribution to the training corpus was extracted to be used for an alternate evaluation of the performance of systems.

Tables 1 and 2 show the performance of different tested models using the full and 1k sets. For the rating of the developed system, 3 different systems have been presented for each subtask. Our submitted models consist in different features as follows:

- **Run 1:** Words and lemmas based polarity dictionaries as features, differing between positive and negative scores and between different datasets. Okapi BM25 scores of mono-grams used as features with the lemmas of the tweet texts. Binary classifiers were implemented using LinearSVC Machines and the global classifier uses their predictions (True or False).
- **Run 2:** Words and lemmas based polarity dictionaries as features, differing between positive and negative scores and between different datasets. Okapi BM25 scores of mono-grams and bi-grams used as features with the lemmas of the tweet texts. Binary classifiers were implemented using LinearSVC Machines and the global classifier uses their predictions (True or False).
- **Run 3:** Similar to **Run 2**, with the exception of the binary classifiers that were implemented using Support Vector Regression (SVR) Machines and the global classifier uses their predictions (0 to 1 float values).

	Run	Accuracy
6 Labels	Run1	**0.560**
	Run2	0.557
	Run3	0.545
4 Labels	Run1	0.608
	Run2	**0.625**
	Run3	0.490

Table 1: Accuracy on the 5 levels and 3 levels of different approaches using the General Corpus.

The systems based on SVM present the best accuracy levels, with an appreciably higher performance in all tests than the system based in SVR. This suggests that the precision of the regression values, in contrast with the binary values of the SVM classifiers, has a negative impact on the global classifier. However, the use of mono-grams and bi-grams as features presents different success rates depending of the test. This part of the system must be analysed in-depth in order to comprehend the performance difference between both systems.

	Run	Accuracy
6 Labels (1k)	Run1	0.407
	Run2	**0.408**
	Run3	0.396
4 Labels (1k)	Run1	**0.601**
	Run2	0.583
	Run3	0.571

Table 2: Accuracy on the 5 levels and 3 levels of different approaches using the 1k Test Corpus.

5 Conclusions and Future work

This paper describes the participation of the DeustoTech Internet research group in the Task 1: Sentiment Analysis at global level at TASS 2015. In our first participation, our team presents a system based in Support Vector Machines in conjunction with several well established polarity lexicons. Experimental results present a good baseline to continue working through the development of new models and developing an structure able to take full advantage of multiple supervised learning systems.

As future work, we propose to research on different approaches to aboard the measure of sentiment analysis problems, especially those related to sentiment degrees with the aim to detect clearly differences between different sentiment levels (Good vs Very Good, for example).

For further work, we would like to improve the present system including some steps previously to the classifier module, that have been demonstrated to improve the final results like a normalization pipeline based on tweets. Also, the necessity of improving the tokenization module to include features like punctuation signs, web addresses, and named entities has become apparent.

Acknowledgments

The research activities described in this paper are funded by DeustoTech INTERNET, Deusto Institute of Technology, a research institute within the University of Deusto.

References

Bermingham, A. and A. F. Smeaton. 2011. On using Twitter to monitor political sentiment and predict election results. In *Bermingham, Adam and Smeaton, Alan F. (2011) On using Twitter to monitor political sentiment and predict election results. In: Sentiment Analysis where AI meets Psychology (SAAIP) Workshop at the International Joint Conference for Natural Language Processing (IJCNLP), 13th November 2011, Chiang Mai, Thailand.*, Chiang Mai, Thailand, November.

Dave, K., S. Lawrence, and D. Pennock. 2003. Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews. In *Proceedings of the 12th International Conference on World Wide Web*, WWW '03, pages 519–528, New York, NY, USA. ACM.

Davidov, D., O. Tsur, and A. Rappoport. 2010. Enhanced sentiment learning using twitter hashtags and smileys. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 241–249. Association for Computational Linguistics.

De Choudhury, M., M. Gamon, S. Counts, and E. Horvitz. 2013. Predicting Depression via Social Media. In *ICWSM*.

Gao, H., J. Hu, C. Wilson, Z. Li, Y. Chen, and B. Y. Zhao. 2010. Detecting and Characterizing Social Spam Campaigns. In *Proceedings of the 10th ACM SIGCOMM Conference on Internet Measurement*, IMC '10, pages 35–47, New York, NY, USA. ACM.

Go, A., R. Bhayani, and L. Huang. 2009. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, 1:12.

Han, B. and T. Baldwin. 2011. Lexical Normalisation of Short Text Messages: Makn Sens a #twitter.

Han, B., P. Cook, and T. Baldwin. 2013. unimelb: Spanish Text Normalisation. In *Tweet-Norm@ SEPLN*, pages 32–36.

Hurtado, LF. and F. Pla. 2014. ELiRF-UPV en TASS 2014: Análisis de sentimientos, detección de tópicos y análisis de sentimientos de aspectos en twitter. *Procesamiento del Lenguaje Natural*.

Jakob, N. and I. Gurevych. 2010. Extracting opinion targets in a single-and cross-domain setting with conditional random fields. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1035–1045. Association for Computational Linguistics.

Jungherr, A. 2013. Tweets and Votes, a Special Relationship: The 2009 Federal Election in Germany. In *Proceedings of the 2Nd Workshop on Politics, Elections and Data*, PLEAD '13, pages 5–14, New York, NY, USA. ACM.

M., Hu and Bing L. 2004. Mining and summarizing customer reviews. *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*.

Martínez-Cámara, E., M. T. Martín-Valdivia, M. L. Molina-Gonzalez, and L. A. Ureña-López. 2013. Bilingual experiments on an opinion comparable corpus. *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*.

Martinez-Camara, E., M. T. Martin-Valdivia, L. A. Ureña-Lopez, and A. Montejo-Raez. 2014. Sentiment analysis in Twitter. *Natural Language Engineering*, 20(01):1–28, January.

Mislove, A., S. Lehmann, YY. Ahn, JP. Onnela, and J. N. Rosenquist. 2011. Understanding the Demographics of Twitter Users. In *ICWSM*.

Padró, L. and E. Stanilovsky. 2012. Freeling 3.0: Towards wider multilinguality.

Pak, A. and P. Paroubek. 2010. Twitter as a Corpus for Sentiment Analysis and Opinion Mining. In *LREC*.

Pang, B. and L. Lee. 2008. Opinion Mining and Sentiment Analysis. *Found. Trends Inf. Retr.*, 2(1-2):1–135, January.

Pennacchiotti, M. and AM. Popescu. 2011. A Machine Learning Approach to Twitter User Classification. In *ICWSM*.

Robertson, S.E., S. Walker, S. Jones, Hancock-Beaulieu M. M., and Gatford M. 1995. Okapi at trec-3. *NIST SPECIAL PUBLICATION SP, 109-109.*

Ruiz, P., 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 at SEPLN 2013.*

San Vicente, I. and X. Saralegi. 2014. Looking for features for supervised tweet polarity classification. *Procesamiento del Lenguaje Natural.*

Saralegi, X. and I. San Vicente. 2013. Elhuyar at TASS 2013. *In Proceedings of "XXIX Congreso de la Sociedad Española de Procesamiento de lenguaje natural". Workshop on Sentiment Analysis at SEPLN (TASS 2013). Madrid. pp. 143-150. ISBN: 978-84-695-8349-4.*

Thelwall, M., K. Buckley, and G. Paltoglou. 2012. Sentiment strength detection for the social web. *Journal of the American Society for Information Science and Technology*, 63(1):163–173, January.

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*, pages 417–424. Association for Computational Linguistics.

Vilares, D., Y. Doval, M.A Alonso, and C. Gómez-Rodríguez. 2014. Lys at TASS 2014: A prototype for extracting and analysing aspects from spanish tweets. *Procesamiento del Lenguaje Natural.*

Villena-Román, J., J. García-Morera, M. A. García-Cumbreras, E. Martínez-Cámara, M. T. Martín-Valdivia, and L. A. Ureña-López. 2015. Overview of TASS 2015.

Wang, X., L. Tokarchuk, F. Cuadrado, and S. Poslad. 2013. Exploiting Hashtags for Adaptive Microblog Crawling. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '13, pages 311–315, New York, NY, USA. ACM.

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.

TASS 2015, septiembre 2015, pp 29-34 recibido 17-07-15 revisado 24-07-15 aceptado 28-07-15

Aspect based Sentiment Analysis of Spanish Tweets

Análisis de Sentimientos de Tweets en Español basado en Aspectos

Oscar Araque, Ignacio Corcuera, Constantino Román, Carlos A. Iglesias y **J. Fernando Sánchez-Rada**
Grupo de Sistemas Inteligentes, Departamento de Ingeniería de Sistemas Telemáticos,
Universidad Politécnica de Madrid (UPM), España
Avenida Complutense, nº 30, 28040 Madrid, España
{oscar.aiborra, ignacio.cplatas, c.romang}@alumnos.upm.es
{cif, jfernando}@dit.upm.es

Resumen: En este artículo se presenta la participación del Grupo de Sistemas Inteligentes (GSI) de la Universidad Politécnica de Madrid (UPM) en el taller de Análisis de Sentimientos centrado en tweets en Español: el TASS2015. Este año se han propuesto dos tareas que hemos abordado con el diseño y desarrollo de un sistema modular adaptable a distintos contextos. Este sistema emplea tecnologías de Procesado de Lenguaje Natural (NLP) así como de aprendizaje automático, dependiento además de tecnologías desarrolladas previamente en nuestro grupo de investigación. En particular, hemos combinado un amplio número de rasgos y léxicos de polaridad para la detección de sentimiento, junto con un algoritmo basado en grafos para la detección de contextos. Los resultados experimentales obtenidos tras la consecución del concurso resultan prometedores.
Palabras clave: Aprendizaje automático, Procesado de lenguaje natural, Análisis de sentimientos, Detección de aspectos

Abstract: This article presents the participation of the Intelligent Systems Group (GSI) at Universidad Politécnica de Madrid (UPM) in the Sentiment Analysis workshop focused in Spanish tweets, TASS2015. This year two challenges have been proposed, which we have addressed with the design and development of a modular system that is adaptable to different contexts. This system employs Natural Language Processing (NLP) and machine-learning technologies, relying also in previously developed technologies in our research group. In particular, we have used a wide number of features and polarity lexicons for sentiment detection. With regards to aspect detection, we have relied on a graph-based algorithm. Once the challenge has come to an end, the experimental results are promising.
Keywords: Machine learning, Natural Language Processing, Sentiment analysis, Aspect detection

1 Introduction

In this article we present our participation for the TASS2015 challenge (Villena-Román et al., 2015a). This work deals with two different tasks, that are described next.

The first task of this challenge, Task 1 (Villena-Román et al., 2015b), consists of determining the global polarity at a message level. Inside this task, there are two evaluations: one in which 6 polarity labels are considered (P+, P, NEU, N, N+, None), and another one with 4 polarity labels considered (P, N, NEU, NONE). P stands for *positive*, while N means *negative* and NEU is *neutral*. The "+" symbol is used for intensification of the polarity. It is considered that NONE means absence of sentiment polarity. This task provides a corpus (Villena-Román et al., 2015b), which contains a total of 68.000 tweets written in Spanish, describing a diversity of subjects.

The second and last task, Task 2 (Villena-Román et al., 2015b), is aimed to detect the sentiment polarity at an aspect level using three labels (P, N and NEU). Within this task, two corpora (Villena-Román et al., 2015b) are provided: SocialTV and STOMPOL corpus. We have restricted ourselves to the SocialTV corpus in this edition. This corpus contains 2.773 tweets captured during the celebration of the *2014 Final of Copa del*

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

rey championship[1]. Along with the corpus a set of aspects which appear in the tweets is given. This list is essentially composed by football players, coaches, teams, referees, and other football-related concepts such as crowd, authorities, match and broadcast.

The complexity presented by the challenge has taken us to develop a modular system, in which each component can work separately. We have developed and experimented with each module independently, and later combine them depending on the Task (1 or 2) we want to solve.

The rest of the paper is organized as follows. First, Section 2 is a review of the research involving sentiment analysis in the Twitter domain. After this, Section 3 briefly describes the general architecture of the developed system. Following that, Section 4 describes the module developed in order to confront the Task 1 of this challenge. After this, Section 5 explains the other modules necessaries to address the Task 2. Finally, Section 6 concludes the paper and presents some conclusions regarding our participation in this challenge, as well as future works.

2 Related Work

Centering the attention in the scope of TASS, many researches have experimented, through the TASS corpora, with different approaches to evaluate the performance of these systems. Vilares et al. (2014) present a system relying in machine learning classification for the tasks of sentiment analysis, and a heuristics based approach for aspect-based sentiment analysis. Another example of classification through machine learning is the work of Hurtado and Pla (2014), in which they utilize Support Vector Machine (SVM) with remarkable results. It is common to incorporate linguistic knowledge to this systems, as proposed by Urizar and Roncal (2013), who also employ lexicons in its work. Balahur and Perea-Ortega (2013) deal with this problem using dictionaries and translated data from English to Spanish, as well as machine-learning techniques. An interesting procedure is performed by Vilares, Alonso, and Gómez-Rodríguez (2013): using semantic information added to psychological knowledge extracted from dictionaries, they combine these features to train a machine learning algorithm. Fernández et al. (2013) employ a ranking algorithm using bigrams and added to this a skipgrams scorer, which allow them to create sentiment lexicons that are able to retain the context of the terms. A different approach is by means of the Word2Vec model, used by Montejo-Ráez, García-Cumbreras, and Díaz-Galiano (2104), in which each word is considered in a 200-dimensional space, without using any lexical or syntactical analysis: this allows them to develop a fairly simple system with reasonable results.

3 System architecture

One of ours main goals is to design and develop an adaptable system which can function in a variety of situations. As we have already mentioned, this has taken us to a system composed of several modules that can work separately. Since the challenge proposes two different tasks (Villena-Román et al., 2015b), we will utilize each module when necessary.

Our system is divided into three modules:

- **Named Entity Recognizer (NER)** module. The NER module detects the entities within a text, and classifies them as one of the possibles entities. In the Section 5 a more detailed description of this module and the set of entities given is presented, as it is used in the Task 2.
- **Aspect and Context detection** module. This module is in charge of detecting the remaining aspects -aspects that are not entities and therefore can not be detected as such- and the contexts of all aspects. In the Section 5 this module is described in greater detail since it is only used for tackling the Task 2.
- **Sentiment Analysis** module. As the name suggests, the goal of this module is to classify the given texts using sentiment polarity labels. This module is based on combining NLP and machine learning techniques and is used in both Task 1 and 2. It is explained in more detail next.

3.1 Sentiment Analysis module

The sentiment analysis module relies in a SVM machine-learning model that is trained with data composed of features extracted from the TASS dataset: General corpus for

[1] www.en.wikipedia.org/wiki/2014_Copa_del_Rey_Final

the Task 1 and SocialTV corpus for Task 2 (Villena-Román et al., 2015b).

3.1.1 Feature Extraction

We have used different approaches to design the feature extraction. The reference document taken in the development of the features extraction was made by Mohammad, Kiritchenko, and Zhu (2013). With this in mind, the features extracted from each tweet to form a feature vector are:

- *N-grams*, combination of contiguous sequences of one, two and three tokens consisting on words, lemmas and stem words. As this information can be difficult to handle due to the huge volume of N-grams that can be formed, we set a minimum frequency of three occurrences to consider the N-gram.
- *All-caps*, the number of words with all characters in upper cases that appears in the tweets.
- *POS information*, the frequency of each part-of-speech tag.
- *Hashtags*, the number of hashtags terms.
- *Punctuation marks*, these marks are frequently used to increase the sentiment of a sentence, specially on the Twitter domain. The presence or absence of these marks (?!) are extracted as a new feature, as well as its relative position within the document.
- *Elongated words*, the number of words that has one character repeated more than two times.
- *Emoticons*, the system uses a Emoticons Sentiment Lexicon, which has been developed by Hogenboom et al. (2013).
- *Lexicon Resources*, for each token w, we used the sentiment score $score(w)$ to determine:
  1. Number of words that have a $score(w) \neq 0$.
  2. Polarity of each word that has a $score(w) \neq 0$.
  3. Total score of all the polarities of the words that have a $score(w) \neq 0$.

  The best way to increase the coverage range with respect to the detection of words with polarity is to combine several resources lexicon. The lexicons used are: Elhuyar Polar Lexicon (Urizar and Roncal, 2013), ISOL (Martínez-Cámara et al., 2013), Sentiment Spanish Lexicon (SSL) (Veronica Perez Rosas, 2012), SOCAL (Taboada et al., 2011) and ML-SentiCON (Cruz et al., 2014).
- *Intensifiers*, a intensifier dictionary (Cruz et al., 2014) has been used for calculating the polarity of a word, increasing or decreasing its value.
- *Negation*, explained in 3.1.2.
- *Global Polarity*, this score is the sum of the punctuations from the emoticon analysis and the lexicon resources.

3.1.2 Negation

An important feature that has been used to develop the classifier is the treatment of the negations. This approach takes into account the role of the negation words or phrases, as they can alter the polarity value of the words or phrases they precede.

The polarity of a word changes if it is included in a negated context. For detecting a negated context we have utilized a set of negated words, which has been manually composed by us. Besides, detecting the context requires deciding how many tokens are affected by the negation. For this, we have followed the proposal by Pang, Lee, and Vaithyanathan (2002).

Once the negated context is defined there are two features affected by this: N-grams and lexicon. The negation feature is added to these features, implying that its negated (e.g. positive becomes negative, +1 becomes -1). This approximation is based on the work by Saurí and Pustejovsky (2012).

4 Task 1: Sentiment analysis at global level

4.1 Experiment and results

In this competition it is allowed for submission up to three experiments for each corpus. With this in mind, three experiments have been developed in this task attending to the lexicons that adjust better to the corpus:

- *RUN-1*, there is one lexicon that is adapted well to the corpus, the ElhPolar lexicon. It has been decided to use only this dictionary in the first run.

- *RUN-2*, in this run the two lexicons that have the best results in the experiments have been combined, the ElhPolar and the ISOL.
- *RUN-3*, the last run is a mix of all the lexicon used on the experiments.

Experiment	Accuracy	F1-Score
6labels	61.8	50.0
6labels-1k	48.7	44.6
4labels	69.0	55.0
4labels-1k	65.8	53.1

Table 1: Results of RUN-1 in the Task 1

Experiment	Accuracy	F1-Score
6labels	61.0	49.5
6labels-1k	48.0	44.0
4labels	67.9	54.6
4labels-1k	64.6	53.1

Table 2: Results of RUN-2 in the Task 1

Experiment	Accuracy	F1-Score
6labels	60.8	49.3
6labels-1k	47.9	43.7
4labels	67.8	54.5
4labels-1k	64.6	48.7

Table 3: Results of RUN-3 in the Task 1

5 Task 2: Aspect-based sentiment analysis

This task is an extension of the Task 1 in which sentiment analysis is made at the aspect level. The goal in this task is to detect the different aspects that can be in a tweet and afterwards analyze the sentiment associated with each aspect.

For this, we used a pipeline that takes the provided corpus as input and produces the sentiment annotated corpus as output. This pipeline can be divided into three major modules that work in a sequential manner: first the NER, second the Aspect and Context detection, and third the Sentiment Analysis as described below.

5.1 NER

The goal of this module is to detect the words that represent a certain entity from the set of entities that can be identified as a *person* (players and coaches) or an *organization* (teams).

For this module we used the Stanford CRF NER (Finkel, Grenager, and Manning, 2005). It includes a Spanish model trained on news data. To adapt the model, we trained it instead with the training dataset (Villena-Román et al., 2015b) and a gazette. The model is trained with two labels: *Person* (PER) and *Organization* (ORG). The gazette entries were collected from the training dataset, resulting in a list of all the ways the entities (players, teams or coaches) were named. We verified the performance of the Stanford NER by means of cross-validation on the training data. With this, we obtained an average F1-Score of 91.05%.

As the goal of the NER module is to detect the words that represent a specific entity, we used a list of all the ways these entities were named. In this way, once the Stanford NER detect the general entity our improved NER module search in this list and decides the particular entity by matching the pattern of the entity words.

5.2 Aspect and Context detection

This module aims to detect the aspects that are not entities, and thus have not been detected by the NER module. To achieve this, we have composed a dictionary using the training dataset (Villena-Román et al., 2015b) which contains all the manners that all the aspects -including the entities formerly detected- are named. Using this dictionary, this module can detect words that are related to a specific aspect. Although the NER module already detects entities as players, coaches or teams, this module can detect them too: it treats these detected entities as more relevant than its own recognitions, combining in this way the capacity of aspect/entity detection of the NER module and this module.

As for the context detection, we have implemented a graph based algorithm (Mukherjee and Bhattacharyya, 2012) that allows us to extract sets of words related to an aspect from a sentence, even if this sentence has different aspects and mixed emotions. The context of an aspect is the set of words related

to that aspect. Besides, we have extended this algorithm in such a way that allow us to configure the scope of this context detection.

Combining this two approaches -aspect and context detection- this module is able to detect the word or words which identify an aspect, and extract the context of this aspect. This context allows us to isolate the sentiment meaning of the aspect, fact that will be very interesting for the sentiment analysis at an aspect level.

We have obtained an accuracy of 93.21% in this second step of the pipeline with the training dataset (Villena-Román et al., 2015b). As for the test dataset (Villena-Román et al., 2015b) we obtained an accuracy of 89.27%[2].

5.3 Sentiment analysis

The sentiment analysis module is the end of the processing pipeline. This module is in charge of classifying the detected aspects in polarity values through the contexts of each aspect. We have used the same model used in Task 1 to analyse every detected aspect in Task 2, given that the detected aspect contexts in Task 2 are similar to the texts analysed in Task 1.

Nevertheless, though using the same model, it is needed to train this model with the proper data. For this, we extracted the aspects and contexts from the train dataset, process the corresponding features (explained in Section 3), and then train the model with these. In this way, the trained machine is fed contexts of aspects that will classify in one of the three labels (as mentioned: positive, negative and neutral).

5.4 Results

By means of connecting these three modules together, we obtain a system that is able to recognize entities and aspects, detect the context in which they are enclosed, and classify them at an aspect level. The performance of this system is showed in the Table 4. The different RUNs represent separate adjustments of the same experiment, in which several parameters are controlled in order to obtain the better performance.

As can be seen in Table 4, the global performance obtained is fairly positive, as our system ranked first in F1-Score and second in Accuracy.

[2]We calculated this metric using the output granted by the TASS uploading page www.daedalus.es/TASS2015/private/evaluate.php.

Experiment	Accuracy	F1-Score
RUN-1	**63.5**	**60.6**
RUN-2	62.1	58.4
RUN-3	55.7	55.8

Table 4: Results of each run in the Task 2

6 Conclusions and future work

In this paper we have described the participation of the GSI in the TASS 2015 challenge (Villena-Román et al., 2015a). Our proposal relies in both NLP and machine-learning techniques, applying them jointly to obtain a satisfactory result in the rankings of the challenge. We have designed and developed a modular system that relies in previous technologies developed in our group (Sánchez-Rada, Iglesias, and Gil, 2015). These characteristics make this system adaptable to different conditions and contexts, feature that results very useful in this competition given the diversity of tasks (Villena-Román et al., 2015b).

As future work, our aim is to improve aspect detection by including semantic similarity based on the available lexical resources in the Linguistic Linked Open Data Cloud. To this aim, we will integrate also vocabularies such as Marl (Westerski, Iglesias, and Tapia, 2011). In addition, we are working on improving the sentiment detection based on the social context of users within the MixedEmotions project.

Acknowledgement

This research has been partially funded and by the EC through the H2020 project MixedEmotions (Grant Agreement no: 141111) and by the Spanish Ministry of Industry, Tourism and Trade through the project Calista (TEC2012-32457). We would like to thank Maite Taboada as well as the rest of researchers for providing us their valuable lexical resources.

References

Balahur, A. and José M. Perea-Ortega. 2013. Experiments using varying sizes and machine translated data for sentiment analysis in twitter.

Cruz, Fermín L, José A Troyano, Beatriz Pontes, and F Javier Ortega. 2014. Building layered, multilingual sentiment lexicons at synset and lemma levels. *Expert Systems with Applications*, 41(13):5984–5994.

Fernández, J., Y. Gutiérrez, J. M. Gómez, P. Martínez-Barco, A. Montoyo, and R. Muñoz. 2013. Sentiment analysis of Spanish tweets using a ranking algorithm and skipgrams.

Finkel, Jenny Rose, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. pages 363–370.

Hogenboom, A., D. Bal, F. Franciscar, M. Bal, F. De Jong, and U. Kaymak. 2013. Exploiting emoticons in polarity classification of text.

Hurtado, Ll. and F. Pla. 2014. ELiRF-UPV en TASS 2014: Análisis de sentimientos, detección de tópicos y análisis de sentimientos de aspectos en twitter.

Martínez-Cámara, E., M. Martın-Valdivia, MD Molina-González, and L. Ureña López. 2013. Bilingual experiments on an opinion comparable corpus. *WASSA 2013*, 87.

Mohammad, Saif M, Svetlana Kiritchenko, and Xiaodan Zhu. 2013. Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets. In *Second Joint Conference on Lexical and Computational Semantics (* SEM)*, volume 2, pages 321–327.

Montejo-Ráez, A, M.A. García-Cumbreras, and M.C. Díaz-Galiano. 2104. Participación de SINAI Word2Vec en TASS 2014.

Mukherjee, Subhabrata and Pushpak Bhattacharyya. 2012. Feature specific sentiment analysis for product reviews. volume 7181 of *Lecture Notes in Computer Science*, pages 475–487. Springer.

Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up?: sentiment classification using machine learning techniques. In *Proc. of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 79–86. Association for Computational Linguistics.

Sánchez-Rada, J. Fernando, Carlos A. Iglesias, and Ronald Gil. 2015. A Linked Data Model for Multimodal Sentiment and Emotion Analysis. 4th Workshop on Linked Data in Linguistics: Resources and Applications.

Saurí, Roser and James Pustejovsky. 2012. Are you sure that this happened? assessing the factuality degree of events in text. *Computational Linguistics*, 38(2):261–299.

Taboada, Maite, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2):267–307.

Urizar, Xabier Saralegi and Iñaki San Vicente Roncal. 2013. Elhuyar at TASS 2013.

Veronica Perez Rosas, Carmen Banea, Rada Mihalcea. 2012. Learning sentiment lexicons in spanish. In *Proc. of the international conference on Language Resources and Evaluation (LREC)*, Istanbul, Turkey.

Vilares, D., M. A. Alonso, and C. Gómez-Rodríguez. 2013. LyS at TASS 2013: Analysing Spanish tweets by means of dependency parsing, semantic-oriented lexicons and psychometric word-properties.

Vilares, David, Yerai Doval, Miguel A. Alonso, and Carlos Gómez-Rodríguez. 2014. LyS at TASS 2014: a prototype for extracting and analysing aspects from Spanish tweets.

Villena-Román, Julio, Janine García-Morera, Miguel A. García-Cumbreras, Eugenio Martínez-Cámara, M. Teresa Martín-Valdivia, and L. Alfonso Ureña-López, editors. 2015a. *Proc. of TASS 2015: Workshop on Sentiment Analysis at SEPLN*, number 1397 in CEUR Workshop Proc., Aachen.

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

Westerski, Adam, Carlos A. Iglesias, and Fernando Tapia. 2011. Linked Opinions: Describing Sentiments on the Structured Web of Data. In *Proc. of the 4th International Workshop Social Data on the Web.*

TASS 2015, septiembre 2015, pp 35-40 recibido 14-07-15 revisado 24-07-15 aceptado 29-07-15

GTI-Gradiant at TASS 2015: A Hybrid Approach for Sentiment Analysis in Twitter*

GTI-Gradiant en TASS 2015: Una aproximación híbrida para el análisis de sentimiento en Twitter

Tamara Álvarez-López
Jonathan Juncal-Martínez
Milagros Fernández-Gavilanes
Enrique Costa-Montenegro
Francisco Javier González-Castaño
GTI Research Group, AtlantTIC
University of Vigo, 36310 Vigo, Spain
{talvarez,jonijm,milagros.fernandez,
kike}@gti.uvigo.es, javier@det.uvigo.es

Hector Cerezo-Costas
Diego Celix-Salgado
Gradiant
36310 Vigo, Spain
{hcerezo,dcelix}@gradiant.org

Resumen: Este artículo describe la participación en el workshop TASS 2015 del grupo de investigación GTI, del centro AtlantTIC, perteneciente a la Universidad de Vigo, y el centro tecnológico Gradiant. Ambos grupos han desarrollado conjuntamente una aproximación híbrida para el análisis de sentimiento global en Twitter, presentado en la tarea 1 del TASS. Se propone un sistema basado en clasificadores y en aproximaciones sin supervisión, construidas mediante léxicos de polaridad y estructuras sintácticas. La combinación de los dos tipos de sistemas ha proporcionado resultados competitivos sobre los conjuntos de prueba propuestos.
Palabras clave: Léxico polar, análisis de sentimiento, dependencias sintácticas.

Abstract: This paper describes the participation of the GTI research group of AtlantTIC, University of Vigo, and Gradiant (Galician Research and Development Centre in Advanced Telecommunications), in the TASS 2015 workshop. Both groups have worked together in the development of a hybrid approach for sentiment analysis, at a global level, of Twitter, proposed in task 1 of TASS. A system based on classifiers and unsupervised approaches, built with polarity lexicons and syntactic structures, is presented here. The combination of both approaches has provided highly competitive results over the given datasets.
Keywords: Polarity lexicon, sentiment analysis, dependency parsing.

1 Introduction

In recent years, research on the field of *Sentiment Analysis* (SA) has increased considerably, due to the growth of user content generated in social networks, blogs and other platforms on the Internet. These are considered valuable information for companies, which seek to know or even predict the acceptance of their products, to design their marketing campaigns more efficiently. One of these sources of information is Twitter, where users are allowed to write about any topic, using colloquial and compact language. As a consecuence, SA in Twitter is specially challenging, as opinions are expressed in one or two short sentences. Moreover, they include special elements such as hashtags or mentions. Henceforth, additional treatments must be applied when analyzing a tweet.

Numerous contributions on this subject can be found in the literature. Most of them are supervised machine learning approaches, although unsupervised semantic can also be found in this field. The first ones are usually classifiers built from features of a "bag of words" representation (Pak and Paroubek, 2010), whilst the second ones try to model linguistic knowledge by using polarity dictionaries (Brooke, Tofiloski, and Taboada,

* This work was supported by the Spanish Government, co-financed by the European Regional Development Fund (ERDF) under project TACTICA, and RedTEIC (R2014/037).

2009), which contain words tagged with their semantic orientation. These strategies involve lexics, syntax or semantics analytics (Quinn et al., 2010) with a final aggregation of their values.

The TASS evaluation workshop aims at providing a benchmark forum for comparing the latest approaches in this field. In this way, our team only took part in task 1 related to SA in Twitter. This task encompasses four experiments. The first consists of evaluating tweet polarities over a big dataset of tweets, with only 4 tags, *positive* (P), *negative* (N), *neutral* (NEU) or *no opinion* (NONE) expressed. In the second experiment, the same evaluation is requested over a smaller selection of tweets. The third and fourth experiments propose the same two datasets, respectively, but with 6 different possible tags, including *strong positive* (P+) and *strong negative* (N+). In addition, a training set has been provided, in order to build the models (Villena-Román et al., 2015).

The rest of this article is structured as follows: Section 2 presents in detail the system proposed. Section 3 describes the results obtained and some experiments performed over the target datasets. Finally, Section 4 summarizes the main findings and conclusions.

2 System overview

Our system is a combination of two different approaches. The first approach is an unsupervised approach, based on sentiment dictionaries, which are automatically generated from the set of tweets to analyze (a set of positive and negative seeds, created manually, are necessary to start the process). The second is a supervised approach, which employs *Conditional Random Fields* (CRFs) (Sutton and McCallum, 2011) to detect the scope of potential polarity shifters (e.g. intensification, reversal verbs and negation particles). This information is combined to conform high-level features which are fed to a statistical classifier to finally obtain the polarity of the message.

In this way, both approaches have been previously adapted to the English language and submitted to the *SemEval-2015* sentiment analysis task, achieving good rankings and results separately (Fernández-Gavilanes et al., 2015; Cerezo-Costas and Celix-Salgado, 2015). Because both have shown particular advantages, we decided to build a hybrid system. The following subsections explain the two approaches, as well as the strategy followed to combine them.

2.1 Previous steps

The first treatments to be applied over the set of tweets rely on *natural language processing* (NLP) and are common to both approaches. They include preprocessing, lexical and syntactic analysis and generation of sentiment lexicons.

2.1.1 Preprocessing

The language used on Twitter contains words that are not found in any dictionary, because of orthographic modifications. The aim here is to normalize the texts to get closer to formal language. The actions executed in this stage are the substitution of emoticons, which are divided in several categories, by equivalent Spanish words, for example, :) is replaced by *e_feliz*; the substitution of frequent abbreviations; the removal of repeated characters and the replacement of specific Twitter words such as hashtags, as well as mentions or URLs, by *hashtag*, *mencion* and URL tags, respectively.

2.1.2 Lexical and syntactic analysis

After the preproccessing, the input text is morphologically tagged to obtain the *part-of-speech* (PoS) associated with a word, as adjectives, adverbs, nouns and verbs. Finally, a dependency tree is created with the syntactic functions annotated. These steps are performed with the *Freeling tool* (Padró and Stanilovsky, 2012).

2.1.3 Sentiment lexicons

Sentiment lexicons have been used in many supervised and unsupervised approaches for sentiment detection. They are not so common in Spanish as in English, although there are some available, such as SOCAL (Brooke, Tofiloski, and Taboada, 2009), SpanishDAL (Dell' Amerlina Ríos and Gravano, 2013) and eSOL lexicon (Molina-González et al., 2013). Some of them are lists of words with an associated number, which represents the polarity, and others are just lists of negative and positive words.

However, these dictionaries are not contextualized, so we generate additional ones automatically from the words in the syntactic dependencies of each tweet, considering verbs, nouns and adjectives. Then, we apply a polarity expansion algorithm based on

graphs (Cruz et al., 2011). The starting point of this algorithm is a set of positive and negative words, used as seeds, extracted from the most negative and positive words in the general lexicons. This dictionary will contain a list of words with their polarity associated, which is a real number in [-5, 5]. Finally, we merge each general lexicon with the automatically created ones, obtaining several dictionaries, depending on the combination applied, to feed our system.

As explained in the next sections, the dictionaries obtained must be adapted for using them in the supervised approach. In this case, only a list of positive and negative words is required, with no associated values.

2.2 Supervised approach

This subsection presents the supervised approach for the tagging of Spanish tweets. After the previous steps, lexical, PoS and CRF labels are jointly combined to build the final features that define the input to a logistic regression classifier.

The system works in two phases. First, a learning phase is applied in which the system learns the parameters of the supervised model using manually tagged data. Second, the supervised model is only trained with the training vector provided by the organization.

2.2.1 Strategy initialization

This strategy uses several dictionaries as an input for different steps of the feature extraction process. Hence, a polarity dictionary, previously created and adapted, containing positive and negative words, is provided as input in this step. Certain polarity shifters play an important role in the detection of the polarity of a sentence. Previous attempts in the academic literature followed different approaches, like hand-crafted rules (Sidorov et al., 2013) or CRFs (Lapponi et al., 2012). We employ CRFs to detect the scope of the polarity shifters such as denial particles (e.g. *sin* (`without`), *no* (`no`)) and reversal verbs, (e.g. *evitar* (`avoid`), *solucionar* (`solve`)). In order to obtain the list of reversal verbs and denial particles, basic syntactic rules and a manual supervision were applied to the final system. A similar approach can be found in Choi and Cardie (2008) .

Additional dictionaries are used in the system (e.g. adversative particles or superlatives) but their main purpose is to give support of the learning steps with the polarity shifters.

2.2.2 Polarity modifiers

Polarity shifters are specific particles (e.g. *no* (`no`)), words (e.g. *evitar* (`avoid`)), or constructions (e.g. *fuera de* (`out of`)) that modify the polarity of the words under their influence. Detecting these scopes of influence closely related to the syntactic graphs is difficult due to the unreliability of dependency and syntactic parsers on Twitter. To solve this problem we trained sequential CRFs for each problem we wanted to solve. CRFs are supervised techniques that assign a label to each component (in our case the words of a sentence).

Our system follows a similar approach to Lapponi, Read and Ovrelid (2012) but it has been enhanced to track intensification, comparisons within a sentence, and the effect of adversative clauses (e.g. sentences with *pero* (`but`) particles). We refer the reader to Cerezo-Costas and Celix-Salgado (2015) to see the input features employed by the CRFs.

2.2.3 Classifier

All the characteristics from previous steps are included as input features of a statistical classifier. The lexical features (word, stem, word and stem bigrams and flags extracted from the polar dictionaries) are included with PoS and the labels from the CRFs. The learning algorithm employed was a logistic regressor. Due to the size of the feature space and its sparsity, l1 (0.000001) and l2 (0.00005) regularization was applied to learn the most important features and discard the least relevant for the task.

2.3 Unsupervised approach

The unsupervised approach is based on generated polarity lexicons applied to the syntactic structures previously obtained. The final sentiment result of each tweet is expressed as a real number, calculated as follows: first, the words in the dependency tree are assigned a polarity from the sentiment dictionary; second, a polarity value propagation based on Caro and Grella (2013) is performed on each dependency tree from the lower nodes to the root, by means of propagation rules explained later. The real end value is classified as P, N, NEU or NONE, according to defined intervals.

2.3.1 Intensification rules

Usually, adverbs act as intensifiers or diminishers of the word that follows them.

For example, there is a difference between *bonito* (`beautiful`) and *muy bonito* (`very beautiful`). The first one has a positive connotation, whose polarity is increased by the adverb *muy* (`very`). So, its semantic orientation is altered. Therefore, the intensification is achieved by assigning a positive or negative percentage in the intensifiers and diminishers (Zhang, Ferrari, and Enjalbert, 2012).

2.3.2 Negation rules

If words that imply denial appear in the text, such as *no* (`no`), *nunca* (`never`) or *ni* (`neither`) (Zhang, Ferrari, and Enjalbert, 2012), the meaning is completely altered. For example, there is a difference between *Yo soy inteligente* (`I am intelligent`) and *Yo no soy inteligente* (`I am not intelligent`). The meaning of the text changes from positive to negative, due to the negator nexus. Therefore, the negation is identified by detecting the affected scope in the dependency tree, for subsequently applying a negative factor to all affected nodes.

2.3.3 Polarity conflict rules

Sometimes, two words appearing together express opposite sentiments. The aim here is to detect these cases, known as *polarity conflicts* (Moilanen and Pulman, 2007). For example, in *fiesta aburrida* (`boring party`), *fiesta* (`party`) has a polarity with a positive connotation, which is reduced by the negative polarity of *aburrida* (`boring`). Moreover, in *náufrago ileso* (`unharmed castaway`), *náufrago* (`castaway`) has a negative polarity, which is reduced by the positive polarity of *ileso* (`unharmed`), yielding a new positive connotation.

2.3.4 Adversative/concessive rules

There is a point in common between adversative and concessive sentences. In both cases, one part of the sentence is in contrast with the other. While the former express an objection in compliance with what is said in the main clause, the latter express a difficulty in fulfilling the main clause. We can assume that both constructions will restrict, exclude, amplify or diminish the sentiment reflected in them. Some adversative nexus can be *pero* (`but`) or *sin embargo* (`however`) (Poria et al., 2014), whereas concessive ones can be *aunque* (`although`) or *a pesar de* (`in spite of`) (Rudolph, 1996). For example, in the adversative sentence *Lo había prometido, pero me ha sido imposible* (`I had promised it, but it has been impossible`), the most important part is the one with the nexus, whereas in the concessive sentence *A pesar de su talento, han sido despedidos* (`In spite of their talent, they have been fired`), it is the part without the nexus.

2.4 Combination strategy: the hybrid approach

In order to decide the final polarity of each tweet, we combine both approaches as follows: applying the supervised approach, 15 different outputs are generated, randomizing the training vector and selecting a subset of them for training (leaving out 1500 records in each iteration). Then, another 15 outputs are generated applying the unsupervised approach, using 15 different lexicons, created by combining each general lexicon (SDAL, SO-CAL, eSOL) with the automatically generated one, and also combining 3 or 4 of them. During this process, when a word appears in several dictionaries, we apply a weighted average, varying the relevance assigned to each dictionary, thus providing more output combinations. Afterwards, we apply a majority voting method among the 30 outputs obtained to decide the final tweet polarity. This strategy has shown better performance than only one of the approaches by itself, making the combination of both a good choice for the experiments, as explained in the next section.

3 Experimental results

The performance in task 1 was measured by means of the accuracy (correct tweets according to the gold standard). Table 1 shows the results, where accuracy is represented for each experiment, as well as the results of the top ranking systems, out of 16 participants for the 6-tag subtasks, and 15 participants for the 4-tag subtasks.

It can be noticed that the results for 6 tags are considerably worse than those for 4 tags. It appears that it becomes more difficult for our system, and for any system in general, to detect positive or negative intensities, rather than just distinguishing positive from negative. Furthermore, we can also observe in the results for the smaller dataset that accuracy diminishes notably for both experiments.

As previously said, in order to obtain our results, we combined both approaches, by means of a majority voting method. On the

Team	Accuracy			
	6	6 (1k)	4	4 (1k)
LIF	67.2_2	51.6_1	72.6_1	69.2_1
GTI-GRAD	59.2_5	50.9_2	69.5_3	67.4_2
ELIRF	67.3_1	48.8_3	72.5_2	64.5_5
GSI	61.8_3	48.7_4	69.0_4	65.8_3
LYS	56.8_6	43.4_5	66.4_5	63.4_9

Table 1: GTI-Gradiant accuracy obtained for each experiment, compared to the top ranking systems. The subscripts represent the position in the ranking.

one hand, the outputs resulting from the supervised approach were generated by applying classifiers, with different training records. On the other hand, the unsupervised approach requires the use of several dictionaries, getting a real number polarity for each tweet, and then applying an interval to determine when a tweet carries an opinion or not. This interval is fixed to [-1, 1] for no opinion. In addition, the number of words containing a polarity is taken into account to decide the neutrality of a tweet. That is, if it contains polar words but the total result lies in [-1, 1], this means that there is a contraposition of opinions, so the tweet is tagged as neutral. However, our combined system seemed to work not so well for neutral texts, specially in the bigger datasets. This may be due to the small proportion of neutral tweets through out the whole dataset, as they only represent a 2.15% of the total number of tweets, rising to 6.3% for the small datasets.

For the 6-tag experiments, P+ and N+ tags were determined with the supervised approach. This decision was taken because the unsupervised approach was not able to discriminate efficiently between P and P+ or between N and N+.

Table 2 shows several experiments with the supervised and unsupervised models, as well as with the combined one, so we can appreciate the improvement in the last case. These results were obtained by applying a majority voting method to each approach separately, with 15 outputs, and then to 30 outputs of the combined result.

Approach	Accuracy			
	6	6 (1k)	4	4 (1k)
Supervised	58.4	48.3	66.4	63.8
Unsupervised	47.8	41.8	66.3	65.1
Combined	59.2	50.9	69.5	67.4

Table 2: Comparative accuracy analysis. Both approaches and combined output.

4 Conclusions

This paper describes the participation of the GTI Research Group (AtlantTIC, University of Vigo) and Gradiant (Galician Research and Development Centre in Advanced Telecommunications) in TASS 2015 Task 1: Sentiment Analysis at global level. We have presented a hybrid system, combining supervised and unsupervised approaches, which has obtained competitive results and a good position in the final ranking.

The unsupervised approach consists of sentiment propagation rules on dependencies, whilst the supervised one is based on classifiers. This combination seems to work considerably well in this task.

There is still margin for improvement, mostly in neutral tweets detection and more refined distinction of degrees of positivity and negativity.

References

Brooke, J., M. Tofiloski, and M. Taboada. 2009. Cross-linguistic sentiment analysis: From English to Spanish. In *Proc. of the Int. Conf. RANLP-2009*, pages 50–54, Borovets, Bulgaria. ACL.

Caro, L. Di and M. Grella. 2013. Sentiment analysis via dependency parsing. *Computer Standards & Interfaces*, 35(5):442–453.

Cerezo-Costas, H. and D. Celix-Salgado. 2015. Gradiant-analytics: Training polarity shifters with CRFs for message level polarity detection. In *Proc. of the 9th Int. Workshop on Semantic Evaluation (SemEval 2015)*, pages 539–544, Denver, Colorado. ACL.

Choi, Y. and C. Cardie. 2008. Learning with Compositional Semantics as Structural Inference for Subsentential Sentiment Analysis. In *Proc. of the Conf. on Empirical Methods in Natural Language Processing*, pages 793–801.

Cruz, F. L., J. A. Troyano, F. J. Ortega, and F. Enríquez. 2011. Automatic expan-

sion of feature-level opinion lexicons. In *Proc. of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*, pages 125–131, Stroudsburg, PA, USA. ACL.

Dell' Amerlina Ríos, M. and A. Gravano. 2013. Spanish dal: A Spanish dictionary of affect in language. In *Proc. of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 21–28, Atlanta, Georgia. ACL.

Fernández-Gavilanes, M., T. Álvarez-López, J. Juncal-Martínez, E. Costa-Montenegro, and F. J. González-Castaño. 2015. GTI: An unsupervised approach for sentiment analysis in Twitter. In *Proc. of the 9th Int. Workshop on Semantic Evaluation (SemEval 2015)*, pages 533–538, Denver, Colorado. ACL.

Lapponi, E., J. Read, and L. Ovrelid. 2012. Representing and Resolving Negation for Sentiment Analysis. In *IEEE 12th Int. Conf. on Data Mining Workshops (ICDMW)*, pages 687–692.

Lapponi, E., E. Velldal, L. Øvrelid, and J. Read. 2012. Uio 2: Sequence-Labeling Negation Using Dependency Features. In *Proc. of the 1st Conf. on Lexical and Computational Semantics*, volume 1, pages 319–327.

Moilanen, K. and S. Pulman. 2007. Sentiment composition. In *Proc. of RANLP 2007*, Borovets, Bulgaria.

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 Syst. Appl.*, 40(18):7250–7257.

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

Pak, A. and P. Paroubek. 2010. Twitter as a corpus for sentiment analysis and opinion mining. In *Proc. of the Int. Conf. on Language Resources and Evaluation, LREC 2010, Valletta, Malta.*

Poria, S., E. Cambria, G. Winterstein, and G. Huang. 2014. Sentic patterns: Dependency-based rules for concept-level sentiment analysis. *Knowledge-Based Systems*, 69(0):45–63.

Quinn, K. M., B. L. Monroe, M. Colaresi, M. H. Crespin, and D. R. Radev. 2010. How to analyze political attention with minimal assumptions and costs. *American Journal of Political Science*, 54(1):209–228.

Rudolph, E. 1996. *Contrast: Adversative and Concessive Relations and Their Expressions in English, German, Spanish, Portuguese on Sentence and Text Level.* Research in text theory. Walter de Gruyter.

Sidorov, G., S. Miranda-Jiménez, F. Viveros-Jiménez, A. Gelbukh, N. Castro-Sánchez, F. Velásquez, I. Díaz-Rangel, S. Suárez-Guerra, A. Treviño, and J. Gordon. 2013. Empirical Study of Machine Learning Based Approach for Opinion Mining in Tweets. In Ildar Batyrshin and Miguel González Mendoza, editors, *Advances in Artificial Intelligence*, volume 7629 of *LNCS*. Springer Berlin Heidelberg, pages 1–14.

Sutton, C. and A. McCallum. 2011. An Introduction to Conditional Random Fields. *Machine Learning*, 4(4):267–373.

Villena-Román, J., J. García-Morera, M. A. García-Cumbreras, E. Martínez-Cámara, M. T. Martín-Valdivia, and L. A. Ureña-López. 2015. Overview of TASS 2015. In *TASS 2015: Workshop on Sentiment Analysis at SEPLN.*

Zhang, L., S. Ferrari, and P. Enjalbert. 2012. Opinion analysis: The effect of negation on polarity and intensity. In Jeremy Jancsary, editor, *Proc. of KONVENS 2012*, pages 282–290. ÖGAI, September. PATHOS 2012 workshop.

TASS 2015, septiembre 2015, pp 41-46 recibido 20-07-15 revisado 24-07-15 aceptado 27-07-15

SINAI-EMMA: Vectores de Palabras para el Análisis de Opiniones en Twitter*

SINAI-EMMA: Vectors of words for Sentiment Analysis in Twitter

Eugenio Martínez Cámara, Miguel Á. García Cumbreras, M. Teresa Martín Valdivia y L. ALfonso Ureña López
Departamento de Informática
Universidad de Jaén, E-23071 - Jaén, España
{emcamara, magc, maite, laurena}@ujaen.es

Resumen: Este artículo describe el sistema de clasificación de polaridad desarrollado por el equipo SINAI-EMMA para la tarea 1 del workshop TASS 2015. Nuestro sistema construye vectores de palabras a partir de la información de opinión de 5 recursos lingüísticos. Los resultados obtenidos nos animan a seguir estudiando el aporte de los vectores de palabras a la tarea de Análisis de Opiniones.
Palabras clave: Análisis de Opiniones, Clasificación de la polaridad, recursos léxicos, Espacio Vectorial Semántico

Abstract: In this work, a polarity classification system is developed for the task 1 of workshop TASS 2015 by the SINAI-EMMA team. Our system takes advantage of 5 linguistic resources for building vectors of words. The results encourage us to continue studying the contribution of vectors of words to Sentiment Analysis.
Keywords: Sentiment Analysis, Polarity Classification, linguistic resources, Vector Space Model of Semantics

1 Introducción

TASS (Taller de Análisis de Sentimientos en la SEPLN) es un workshop satélite del congreso de la SEPLN (Sociedad Española para el Procesamiento del Lenguaje Natural), que promueve desde 2012 la evaluación experimental de sistemas de Análisis de Opiniones (AO) sobre *tweets* escritos en español.

Nuestra participación se ha limitado a la tarea 1, denominada *sentiment analysis at global level*, que tiene como objetivo la clasificación de la polaridad de un conjunto de *tweets*. La descripción de la tarea se encuentra en el resumen de la edición del año 2015 de TASS (Villena-Román et al., 2015). El sistema de este año se centra en la subtarea de de clasificación en 6 niveles de polaridad (P+, P, NEU, N, N+, NONE).

La solución que presentamos pasa por extraer características léxicas a partir de n-gramas de cada *tweet*, concretamente *unigramas* y *bigramas*, siendo estas características la polaridad de los n-gramas en distintos recursos léxicos etiquetados. Una vez obtenida la matriz de *tweets* y características utilizamos un sistema de entrenamiento para obtener un modelo léxico, y dicho modelo es el utilizado en la evaluación del subconjunto de test a etiquetar.

El resto del artículo está organizado de la siguiente forma. La siguiente sección describe los modelos de n-gramas y otros modelos del lenguaje aplicados a la detección de polaridad con *tweets*, así como trabajos relacionados. En la Sección 3 se describe el sistema que hemos desarrollado y en la Sección 4 los experimentos realizados, resultados obtenidos y análisis de los mismos. Por último, en la Sección 5 exponemos las conclusiones y el trabajo a realizar.

2 Trabajos relacionados

El Modelo de Espacio Vectorial (MEV) ha demostrado sobradamente su valía para medir la similitud entre documentos, y una muestra es su exitosa aplicación en los sistemas de recuperación de información (Manning, Raghavan, y Schütze, 2008). Por consiguiente, cabe preguntarse si es posible aplicar el mismo esquema para medir la similitud existente entre palabras. Esa pregunta la contestaron afirmativamente Deerwester et al. (1990) al proponer un MEV en el

* Este trabajo ha sido financiado parcialmente por el Fondo Europeo de Desarrollo Regional (FEDER), el proyecto ATTOS del Gobierno de España (TIN2012-38536-C03-0) y el proyecto AORESCU de la Junta de Andalucía (P11-TIC-7684 MO).

que cada palabra estaba caracterizada por un vector, el cual representaba a la palabra en función a su coocurrencia con otras. Al igual que el MEV sobre documentos, el MEV sobre palabras también ha reportado buenos resultados en desambiguación, reconocimiento de entidades, análisis morfológico (*part of speech tagging*) y recuperación de información (Turney y Pantel, 2010; Collobert y Weston, 2008; Turian, Ratinov, y Bengio, 2010).

En la bibliografía relacionada con AO existen trabajos que postulan que las palabras (*unigramas*) son más efectivos para representar la información de opinión (Pang, Lee, y Vaithyanathan, 2002; Martínez-Cámara et al., 2011), y otros en los que se prefiere el uso de n-gramas (Dave, Lawrence, y Pennock, 2003; Kouloumpis, Wilson, y Moore, 2011). La orientación de la opinión de un documento es muy probable que no sea la suma de las polaridades de las palabras individuales que lo componen, sino la combinación de las expresiones que aparecen en el mismo. La palabra "bueno" tiene un significado positivo, pero la expresión "no muy bueno" transmite un significado negativo. Por tanto, en el caso del AO puede que sea más recomendable utilizar n-gramas como unidad de representación de información.

En el presente trabajo tratamos de incrustar al MEV tradicional, una representación vectorial de cada n-grama, dando lugar a un MEV semántico (Turney y Pantel, 2010). Este tipo de modelo tiene una mayor capacidad de representación de la información subyacente en la opinión, al combinar diferentes grados de n-gramas, así como de caracterizar cada n-grama por un conjunto de características. Éste es el fundamento de los nuevos métodos que se están desarrollando para la clasificación de la polaridad. Uno de los primeros trabajos de este nuevo enfoque es (Maas et al., 2011), en el que se consigue que palabras con un significado similar estén representadas por vectores similares. El proceso de generación de los vectores es no supervisado, pero para una correcta clasificación de la polaridad, el método que proponen los autores requiere de información de opinión, de manera que tienen que utilizar un corpus con etiquetas de opinión para conseguir que los vectores asociados a las palabras representen la orientación semántica de la misma.

Socher et al. (2011) tratan de representar la opinión de un conjunto de documentos a través de un MEV a nivel de n-grama, siguiendo en este caso, un modelo de autocodificadores recursivos (*recursive autoencoders*). Bespalov et al. (2012) tratan tener en cuenta en el MEV semántico que diseñan la posición de cada n-grama en el documento, dado que sostienen que la polaridad de un documento depende de la posición en la que se encuentren los n-gramas. El trabajo (Tang et al., 2014) es otro ejemplo de representación de n-gramas por medio de vectores, tratando los autores de insertar información de opinión en dichos vectores mediante el uso de tres redes neuronales. A diferencia de los trabajos anteriores, el de Tang et al. (2014) se centra en la clasificación de la polaridad de *tweets*.

3 Sistema

El sistema que se ha desarrollado para la cuarta edición de TASS trata de trasladar el concepto de MEV semántico a la clasificación de la polaridad de *tweets* en español. En nuestro caso, no se ha pretendido representar a cada palabra en función de su coocurrencia con otras, o teniendo en cuenta su posición en el texto, sino por su valor de polaridad en varios recursos léxicos de opinión: iSOL (Molina-González et al., 2013), SentiWordNet (Baccianella, Esuli, y Sebastiani, 2010), Q-WordNet (Agerri y García-Serrano, 2010), SEL (Rangel, Sidorov, y Suárez-Guerra, 2014) y ML-Senticon (Cruz et al., 2014). Con este esquema de representación, lo que se está haciendo es incrustar en el MEV los distintos puntos de vista que tienen diferentes recursos léxicos de opinión sobre una misma palabra, enriqueciendo de esta forma la representación de la información de opinión que contienen cada una de las palabras.

A continuación se van a describir someramente los recursos lingüísticos que se han utilizado, así como se detallará el proceso de generación de los vectores.

3.1 iSOL

iSOL es una lista de palabras indicadoras de opinión en español. La lista está formada por 8135 palabras, de las cuales 5626 son negativas y 2509 son positivas.

Al tratarse de una lista de palabras de opinión, la única información que aporta es si un término es positivo o negativo, por lo que proporciona dos características binarias al vector de cada palabra.

3.2 SentiWordNet

SentiWordNet es un lista de sentidos de opinión que se construyó siguiendo un enfoque basado en diccionario (Liu, 2012). El diccionario que se empleó en su desarrollo fue WordNet (Miller, 1995). Por tanto, SentiWordNet asocia a cada sentido (*synset*) de WordNet tres valores probabilidad de pertenencia a tres niveles de polaridad: Positivo, Negativo y Neutro.

Cada palabra puede tener asociada varios *synsets*, por lo que se necesita de una función de agregación para obtener un valor de polaridad único para cada uno de los tres niveles de opinión. Al igual que Denecke (2008), se empleó como función de agregación la media aritmética. Por tanto, se calculó la media aritmética de cada nivel de polaridad, obteniéndose de esta manera 3 características, cada una correspondiente a un nivel de polaridad.

3.3 Q-WordNet

Q-WordNet, al igual que SentiWordNet, está basado en WordNet para construir un recurso para la tarea de AO. Los autores de Q-WordNet, al contrario que los de SentiWordNet, consideran la polaridad como una propiedad cualitativa, de manera que una palabra sólo puede ser positiva o negativa. Por consiguiente, Q-WordNet es un recurso de opinión que asocia a cada *synset* de WordNet el valor Positivo o Negativo. El uso de Q-WordNet permite la adición de 2 nuevas características al vector asociado a cada palabra.

3.4 SEL

SEL es una lista de 2036 palabras clasificadas en 6 estados de ánimo diferentes: alegría, enfado, miedo, tristeza, sorpresa y disgusto. Las palabras tienen asociado un valor de probabilidad, que los autores llaman PFA, de pertenencia a una de las 6 categorías.

Para su integración en el sistema, se han transformado las 6 categorías emocionales en dos valores de polaridad, de forma que las categorías alegría y sorpresa se han tomado como positivas, y las clases enfado, miedo, tristeza y disgusto como negativas. Además, se aplicó un filtro a la lista de palabra con el fin de sólo utilizar aquellas que tuvieran un valor de PFA superior a 0,2. Por tanto, tras convertirse SEL a una lista binaria de opinión, se pudieraon generar dos nuevas características binarias de polaridad.

3.5 ML-SentiCon

ML-SentiCon es un recurso de opinión a nivel de lema. Los 26495 que conforman ML-SentiCon se encuentran estratificados en 8 capas. Cada una de las capas representa un nivel de exactitud de pertenencia a la clase positivo o negativo, de manera que existe una mayor probabilidad, de que el significado de opinión que transmiten los lemas de la capa 0 coincida con la clase de opinión que se le ha asignado que los de la capa 7. Al estar los lemas catalogados como positivos o negativos, este recurso también genera dos nuevas características, pero en lugar de ser binarias, su valor se corresponde con la puntuación de polaridad que tiene cada lema en el recurso. En el caso de que el lema sea negativo, su puntuación se multiplica por -1.

4 Clasificación

La preparación de la clasificación consistió en el procesamiento adecuado de los *tweets* para la generación del MEV semántico de *unigramas* y bigramas. La preparación comenzó con la *tokenización* de los *tweets*; aplicación de un proceso de normalización léxica que consistió principalmente en la corrección ortográfica de los tokens; lematización y aplicación de un análisis morfológico.

A continuación se construyeron los vectores de características asociados a cada *unigrama*. En función de la naturaleza de cada recurso, la búsqueda fue distinta. En el caso de iSOL se utilizó como consulta la forma original de cada *unigrama*, a excepción de los *unigramas* que se corresponden con verbos, con los que se usó su lema, es decir, el infinitivo del verbo.

Para SentiWordNet y Q-WordNet la consulta no fue directa porque ambos son dos recursos para inglés. Por tanto, se precisó emplear un recurso que permitiera a partir de palabras en español obtener el *synset* que le corresponde en WordNet. Ese recurso fue la versión en español de WordNet de Multilingual Central Repository (MCR) (Atserias et al., 2004). Una vez que se tenían de MCR los *synsets* asociados, se consultaba SentiWordNet y Q-WordNet para obtener sus correspondientes valores de polaridad.

SEL y ML-Senticon tuvieron un tratamiento similar, dado que la consulta se realizaba utilizando el lema de cada *unigrama*.

Antes de continuar con la descripción de la generación de los *bigramas*, debe precisarse que únicamente se construyeron los vectores de los *unigramas* que al menos estaban recogidos en uno de los cinco recursos lingüísticos que se han considerado en la experimentación. Tras la eliminación de los *unigramas* que no transmiten opinión, el texto del *tweet* se quedaba reducido a dichos *unigramas*. Con esos *unigramas* se desarrolló un proceso de generación de *bigramas*, cuyos vectores se corresponden con la suma vectorial de los vectores de los *unigramas* que constituyen el *bigrama*. La longitud de los vectores de características de opinión de tanto los *unigramas* como de los *bigramas* es de 11 características.

Una vez generado el MEV semántico, es momento de generar el modelo de entrenamiento. Para ello se eligió el algoritmo SVM con *kernel* lineal, ya que en experimentaciones previas ha demostrado su valía para la tarea de AO sobre *tweets*. La implementación que se utilizó fue la de la librería scikit-learn[1] de Python.

5 Resultados

Primeramente, se desarrolló una evaluación con el conjunto de entrenamiento, para evaluar si el esquema descrito anteriormente podría tener éxito.

La evaluación consistió en aplicar un marco de trabajo basado en validación cruzada de 10 particiones (*ten fold cross-validation*). Las medidas de evaluación que se utilizaron fueron la comunes en tareas de clasificación de textos, es decir, Precisión, *Recall*, F1 y *Accuracy*. La tarea a la que se enfrenta el sistema es la de clasificación en seis niveles de de intensidad de opinión, de manera que el uso de la definición clásica de Precisión, *Recall* y F1 no es correcto, debido principalmente a que el resultado estaría sesgado por la clase predominante del conjunto de datos que se está utilizando. Por consiguiente, se decidió usar sus correspondientes macromedidas (*macro-averings*).

Se llevaron a cabo dos evaluaciones, que se diferencian en la manera de aprovechar la información de opinión que aporta el recurso iSOL. En una primera ejecución (SINAI-EMMA_iSOLOriginal), solamente se buscaban en iSOL las formas originales de las palabras convertidas en minúsculas. Si se comprueba el léxico recogido en iSOL se puede comprobar que en él se recogen vocablos, y que uno de los puntales de su éxito es que incluye las derivaciones de género y número propias del español. En cambio, en lo que se refiere a los verbos, no incluye las conjugaciones de los mismos, estando sólo presente el infinitivo. Por tanto, en esta primera ejecución se estaba perdiendo la cobertura de las distintas formas verbales, ya que sólo se podían reconocer aquellas que en el *tweet* aparecieran en infinitivo.

La segunda configuración (SINAI-EMMA_iSOLLema) tuvo en cuenta esa peculiaridad de iSOL, y en el caso de que el término a buscar fuera un verbo, se empleó su lema, es decir su infinitivo. De esta manera se consiguió que un mayor número de vocablos fueran considerados como portadores de opinión. En la Tabla 1 se muestran los resultados alcanzados por estas dos configuraciones.

No es difícil dar una explicación a la ligera mejoría de la configuración SINAI-EMMA_iSOLLema, dado que posibilita aumentar la cobertura del lenguaje por parte de iSOL. Por tanto, la configuración que se utilizó para la clasificación de los *tweets* del subcorpus de test es SINAI-EMMA_iSOLLema. La Tabla 2 recoge los resultados oficiales obtenidos por el sistema, así como el mejor resultado y la media de los resultados del resto de sistemas presentados.

Los resultados evidencian que el sistema se encuentra en la media de los presentados en la edición 2015 de TASS, lo cual por un lado es satisfactorio, ya que el estudio que se ha iniciado en el ámbito de los MEV semánticos reporta, por ahora, unos resultados similares al resto de enfoques que se han presentado, pero por otro, pone de manifiesto que todavía quedan muchos elementos que analizar para seguir avanzando en la resolución del problema de la clasificación de la polaridad de *tweets* en español.

La organización del taller proporciona también una evaluación con un conjunto reducido de 1000 *tweets* etiquetados a mano. La Tabla 3 recoge los resultados obtenidos sobre ese conjunto de datos.

Como se puede apreciar en la Tabla 3, las diferencias entre el sistema presentado y el mejor se acortan, a su vez que el sistema consigue unos resultados superiores a la media según la Macro-Precisión, el Macro-*Recall*

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

Configuración	**Macro-P**	**Macro-R**	**Macro-F1**	***Accuracy***
SINAI-EMMA_iSOLOriginal	36,55 %	36,58 %	35,99 %	40,91 %
SINAI-EMMA_iSOLLema	36,83 %	36,74 %	36,02 %	41,31 %

Tabla 1: Evaluación del sistema de clasificación con el conjunto de entrenamiento.

Configuración	**Macro-P**	**Macro-R**	**Macro-F1**	***Accuracy***
SINAI-EMMA	40,4 %	45,8 %	43,3 %	50,02 %
Mejor	53,1 %	46,5 %	49,6 %	67,3 %
Media	40,9 %	42,2 %	40,9 %	52,7 %

Tabla 2: Resultados de la evaluación oficial.

y el Macro-F1. Ésto indica que el sistema SINAI-EMMA tiene un comportamiento más estable que el resto, porque su rendimiento no experimenta un variación muy acusada ante la modificación del etiquetado del conjunto de evaluación, lo cual es un punto positivo en nuestra investigación.

6 Conclusiones y trabajo a realizar

La principal conclusión a la que se ha llegado es la idoneidad de la representación de la información de opinión mediante un MEV semántico. Además, características de los *unigramas* y *bigramas* se han construido a partir de recursos lingüísticos de opinión. Esto es una primera tentativa a la aplicación de MEV semánticos a la tarea de AO, y los resultados obtenidos nos animan a seguir investigando en esta línea.

El trabajo futuro va a estar dirigido a mejorar el esquema de representación de información, así como en aumentar la información de opinión que representa a cada palabra. Para ello, se va a realizar un estudio del nivel de cobertura de vocabulario que tiene el esquema de representación; se va a tratar de introducir información de la distribución de los *unigramas* y *bigramas* en el corpus por medio del uso de la frecuencia de los mismos en cada documento y en el corpus en general; se va a estudiar la incorporación de un mayor número de recursos de opinión; se va a analizar el aporte para el AO de la incrustación de información de coocurrencia de cada *unigrama* y *bigrama* tomando como referencia un corpus representativo del español; y por último se va a estudiar la utilización de información sintáctica mediante la consideración del efecto de la negación y de los intensificadores.

Bibliografía

Agerri, Rodrigo y Ana García-Serrano. 2010. Q-Wordnet: Extracting polarity from wordnet senses. En *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta; ELRA, may. European Language Resources Association. 19-21.

Atserias, J., L. Villarejo, G. Rigau, E. Agirre, J. Carroll, B. Magnini, y P. Vossen. 2004. The meaning multilingual central repository. En *GWC 2012 6th International Global Wordnet Conference*. Brno: Masaryk University.

Baccianella, S., A. Esuli, y F. Sebastiani. 2010. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. En *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, páginas 2200–2204, Valletta, Malta.

Bespalov, Dmitriy, Yanjun Qi, Bing Bai, y Ali Shokoufandeh. 2012. Sentiment classification with supervised sequence embedding. En *Machine Learning and Knowledge Discovery in Databases*, volumen 7523 de *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, páginas 159–174.

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.

Cruz, F. L., J. A. Troyano, B. Pontes, y F. J. Ortega. 2014. Ml-senticon: Un lexicón

Configuración	**Macro-P**	**Macro-R**	**Macro-F1**	***Accuracy***
SINAI-EMMA	36,6 %	38,0 %	37,3 %	41,1 %
Mejor	44,1 %	45,9 %	45,0 %	51,6 %
Media	36,3 %	37,28 %	36,68 %	41.35 %

Tabla 3: Resultados oficiales sobre el subcorpus de 1000 *tweets*.

multilingüe de polaridades semánticas a nivel de lemas. *Procesamiento del Lenguaje Natural*, 53(0):113–120.

Dave, K., S. Lawrence, y D. M. Pennock. 2003. Mining the peanut gallery: opinion extraction and semantic classification of product reviews. En *Proceedings of the 12th international conference on World Wide Web*, WWW '03, páginas 519–528, New York, NY, USA. ACM.

Deerwester, S., S. T. Dumais, G. W. Furnas, T. K. Landauer, y R. Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

Kouloumpis, Efthymios, Theresa Wilson, y Johanna Moore. 2011. Twitter sentiment analysis: The good the bad and the omg!

Liu, B. 2012. Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, 5(1):1–167.

Maas, A. L., R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, y C. Potts. 2011. Learning word vectors for sentiment analysis. En *Proceedings of the 49th Annual Meeting of the ACL: Human Language Technologies*, volumen 1 de *HLT '11*, páginas 142–150. ACL.

Manning, Christopher D., Prabhakar Raghavan, y Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA.

Martínez-Cámara, Eugenio, M. Teresa Martín-Valdivia, José M. Perea-Ortega, y L. Alfonso Ureña-López. 2011. Técnicas de clasificación de opiniones aplicadas a un corpus en español. *Procesamiento del Lenguaje Natural*, 47:163–170.

Miller, George A. 1995. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41, Noviembre.

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

Pang, Bo, Lillian Lee, y Shivakumar Vaithyanathan. 2002. Thumbs up?: Sentiment classification using machine learning techniques. En *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing*, volumen 10 de *EMNLP '02*, páginas 79–86. ACL.

Rangel, I. D., G. Sidorov, y S. Suárez-Guerra. 2014. Creación y evaluación de un diccionario marcado con emociones y ponderado para el español. *Onomázein*, 1(29):31–46.

Socher, R., J. Pennington, E. H. Huang, A. Y. Ng, y C. D. Manning. 2011. Semi-supervised recursive autoencoders for predicting sentiment distributions. En *Proceedings of the Conference on EMNLP*, EMNLP '11, páginas 151–161. ACL.

Tang, Duyu, Furu Wei, Nan Yang, Ming Zhou, Ting Liu, y Bing Qin. 2014. Learning sentiment-specific word embedding for twitter sentiment classification. En *Proceedings of the 52nd Annual Meeting of the ACL*, volumen 1, páginas 1555–1565, Baltimore, Maryland, June. ACL.

Turian, Joseph, Lev Ratinov, y Yoshua Bengio. 2010. Word representations: A simple and general method for semi-supervised learning. En *Proceedings of the 48th Annual Meeting of the ACL*, ACL '10, páginas 384–394. ACL.

Turney, Peter D. y Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *J. Artif. Int. Res.*, 37(1):141–188, Enero.

Villena-Román, Julio, Janine García-Morera, Miguel A. García-Cumbreras, Eugenio Martínez-Cámara, M. Teresa Martín-Valdivia, y L. Alfonso Ureña López. 2015. Overview of tass 2015. En *TASS 2015: Workshop on Sentiment Analysis at SEPLN*.

TASS 2015, septiembre 2015, pp 47-52 recibido 10-07-15 revisado 24-07-15 aceptado 26-07-15

LyS at TASS 2015: Deep Learning Experiments for Sentiment Analysis on Spanish Tweets*

LyS en TASS 2015: Experimentos con Deep Learning para Análisis del Sentimiento sobre Tweets en Español

David Vilares, Yerai Doval, Miguel A. Alonso and Carlos Gómez-Rodríguez
Grupo LyS, Departamento de Computación, Campus de A Coruña s/n
Universidade da Coruña, 15071, A Coruña, Spain
{david.vilares, yerai.doval, miguel.alonso, carlos.gomez}@udc.es

Resumen: Este artículo describe la participación del grupo LyS en el TASS 2015. En la edición de este año, hemos utilizado una red neuronal denominada *long short-term memory* para abordar los dos retos propuestos: (1) análisis del sentimiento a nivel global y (2) análisis del sentimiento a nivel de aspectos sobre tuits futbolísticos y de política. El rendimiento obtenido por esta red de aprendizaje profundo es comparado con el de nuestro sistema del año pasado, una regresión logística con una regularización cuadrática. Los resultados experimentales muestran que es necesario incluir estrategias como pre-entrenamiento no supervisado, técnicas específicas para representar palabras como vectores o modificar la arquitectura actual para alcanzar resultados acordes con el estado del arte.
Palabras clave: deep learning, long short-term memory, análisis del sentimiento, Twitter

Abstract: This paper describes the participation of the LyS group at TASS 2015. In this year's edition, we used a long short-term memory neural network to address the two proposed challenges: (1) sentiment analysis at a global level and (2) aspect-based sentiment analysis on football and political tweets. The performance of this deep learning approach is compared to our last-year model, based on a square-regularized logistic regression. Experimental results show that strategies such as unsupervised pre-training, sentiment-specific word embedding or modifying the current architecture might be needed to achieve state-of-the-art results.
Keywords: deep learning, long short-term memory, sentiment analysis, Twitter

1 Introduction

The 4th edition of the TASS workshop addresses two of the most popular tasks on *sentiment analysis* (SA), focusing on Spanish tweets: (1) polarity classification at a global level and (2) a simplified version of aspect-based sentiment analysis, where the goal is to predict the polarity of a set of predefined and identified aspects (Villena-Román et al., b).

The challenge of polarity classification has been typically tackled from two different angles: lexicon-based and machine learning (ML) approaches. The first group relies on sentiment dictionaries to detect the subjective words or phrases of the text, and defines lexical- (Brooke, Tofiloski, and Taboada, 2009; Thelwall et al., 2010) or syntactic-based rules (Vilares, Alonso, and Gómez-Rodríguez, 2015c) to deal with phenomena such as negation, intensification or irrealis.

The second group focuses on training classifiers through supervised learning algorithms that are fed a number of features (Pang, Lee, and Vaithyanathan, 2002; Mohammad, Kiritchenko, and Zhu, 2013; Hurtado and Pla, 2014). Although competitive when labelled data is provided, they have shown weakness when interpreting the compositionality of complex phrases (e.g. adversative subordinate clauses). In this respect, some studies have evaluated the impact of syntactic-based features on these supervised learning techniques (Vilares, Alonso, and Gómez-Rodríguez, 2015b; Joshi and Penstein-Rosé, 2009) or other related tasks, such as multi-topic detection on tweets (Vilares, Alonso,

* This research is supported by the Ministerio de Economía y Competitividad y FEDER (FFI2014-51978-C2) and Xunta de Galicia (R2014/034). The first author is funded by the Ministerio de Educación, Cultura y Deporte (FPU13/01180).

and Gómez-Rodríguez, 2015a).

More recently, deep learning (Bengio, 2009) has shown its competitiveness on polarity classification. Bespalov et al. (2011) introduce a word-embedding approach for higher-order n-grams, using a multi-layer perceptron and a linear function as the output layer. Socher et al. (2013) introduce a new deep learning architecture, a Recursive Neural Tensor Network, which improved the state of the art on the Pang and Lee (2005) movie reviews corpus, when trained together with the Stanford Sentiment Treebank. Tang et al. (2014) suggest that currently existing word embedding methods are not adequate for SA, because words with completely different sentiment might appear in similar contexts (e.g. *'good'* and *'bad'*). They pose an sentiment-specific words embedding (SSWE) model, using a deep learning architecture trained from massive distant-supervised tweets. For Spanish, Montejo-Raéz, García-Cumbreras, and Díaz-Galiano (2014) apply word embedding using Word2Vec (Mikolov et al., 2013), to then use those vectors as features for traditional machine learning techniques.

In this paper we also rely on a deep learning architecture, a long short-term memory (LSTM) recurrent neural network, to solve the challenges of this TASS edition. The results are compared with respect to our model for last year's edition, a logistic regression approach fed with hand-crafted features.

2 Task1: Sentiment Analysis at a global level

Let $L=\{l_0, l_1, ..., l_n\}$ be the set of polarity labels and $T=\{t_0, t_1, ..., t_m\}$ the set of labelled texts, the aim of the task consists of defining an hypothesis function, $h : T \rightarrow L$.

To train and evaluate the task, the collection from TASS-2014 (Villena-Román et al., 2015) was used. It contains a training set of 7 128 tweets, intended to build and tune the models, and two test sets: (1) a pooling-labelled collection of 60 798 tweets and (2) a manually-labelled test set of 1 000 tweets. The collection is annotated using two different criteria. The first one considers a set of 6 polarities (L_6): *no opinion* (NONE), *positive* (P), *strongly positive* (P+), *negative* (N), *strongly negative* (N+) and *mixed* (NEU), that are tweets that mix both negative and positive ideas. A simplified version with 4 classes (L_4) is also proposed, where the polarities P+ and N+ are included into P and N, respectively.

In the rest of the paper, we will use h_4 and h_6 to refer our prediction models for 4 and 6 classes, respectively.

3 Task2: Sentiment Analysis at the aspect level

Let $L=\{l_0, l_1, ..., l_n\}$ be the set of polarity labels, $A=\{a_0, a_1, ...a_o\}$ the set of aspects and a $T=\{t_0, t_1, ..., t_m\}$ the set of texts, the aim of the task consists of defining an hypothesis function, $h : A \times T \rightarrow L$. Two different corpora are provided to evaluate this task: a SOCIAL-TV corpus with football tweets (1 773 training and 1 000 test tweets) and a political corpus (784 training and 500 test tweets), called STOMPOL. Each aspect can be assigned the P, N or NEU polarities (L_3).

The TASS organisation provided both A and the identification of the aspects that appear in each tweet, so the task can be seen as identifying the scope $s(a, t)$ of an aspect a in the tweet $t \in T$, with s a substring of t and $a \in A$, to then predict the polarity using the hypothesis function, $h_3(s) \rightarrow L_3$.

To identify the scope we followed a naïve approach: given an aspect a that appears at position i in a text, $t=[w_0, ..., w_{i-x}, ..., a_i, ..., w_{i+x}, ..., w_p]$, we created a snippet of length x that is considered to be the scope of the aspect. Preliminary experiments on the SOCIAL-TV and the STOMPOL corpus showed that $x = 4$ and taking the entire tweet were the best options for each collection, respectively.

4 Supervised sentiment analysis models

Our aim this year was to compare our last-year model to a deep learning architecture that was initially available for binary polarity classification.

4.1 Long Short-Term Memory

Long Short-Term Memory (LSTM) is a recurrent neural network (RNN) proposed by Hochreiter and Schmidhuber (1997). Traditional RNN were born with the objective of being able to store representations of inputs in form of activations, showing temporal capacities and helping to learn short-term dependencies. However, they might suffer from

the problem of *exploding gradients*[1]. The LSTM tries to solve these problems using a different type of units, called *memory cells*, which can remember a value for an arbitrary period of time.

In this work, we use a model composed of a single LSTM and a logistic function as the output layer, which has an available implementation[2] in Theano (Bastien et al., 2012).

To train the model, the tweets were tokenised (Gimpel et al., 2011), lemmatised (Taulé, Martí, and Recasens, 2008), converted to lowercase to reduce sparsity and finally indexed. To train the LSTM-RNN, we relied on ADADELTA (Zeiler, 2012), an adaptive learning rate method, using stochastic training (batch size = 16) to speed up the learning process. Experiments with non-stochastic training runs did not show an improvement in terms of accuracy. We empirically explored the size of the word embedding[3] and the number of words to keep in the vocabulary[4], obtaining the best performance using a choice of 128 and 10 000 respectively.

4.2 L2 logistic regression

Our last-year edition model relied on the simple and well-known squared-regularised logistic regression (L2-LG), that performed very competitively for all polarity classification tasks. A detailed description of this model can be found in Vilares et al. (2014a), but here we just list the features that were used: *lemmas* (Taulé, Martí, and Recasens, 2008), *psychometric properties* (Pennebaker, Francis, and Booth, 2001) and *subjective lexicons* (Saralegi and San Vicente, 2013). This architecture also obtained robust and competitive performance for English tweets, on SemEval 2014 (Vilares et al., 2014b).

Penalising neutral tweets

Previous editions of TASS have shown that the performance on NEU tweets is much lower than for the rest of the classes (Villena-Román et al., a). This year we proposed a small variation on our L2-LG model: a penalising system for NEU tweets to determine the polarities under the L_6 configuration, where: given an L_4 and an L_6 LG-classifier and a tweet t, if $h_6(t) = \text{NEU}$ and $h_4(t) \neq \text{NEU}$ then $h_6(t) := h_4(t)$. The results obtained on the test set shown that we obtained an improvement of 1 percentage point with this strategy (from 55.2% to 56.8% that is reported in the Experiments section).

5 *Experimental results*

Table 1 compares our models with the best performing run of the rest of the participants (out of date runs are not included). The performance of our current deep learning model is still far from the top ranking systems, and from our last-year model too, although it worked acceptably under the L_6 manually-labelled test.

Table 2 and 3 show the F1 score for each polarity, for the LSTM-RNN and L2-LG models, respectively. The results reflect the lack of capacity of the current LSTM model to learn the minority classes in the training data (P, N+ and NEU). In this respect, we plan to explore how balanced corpora and bigger corpora can help diminish this problem.

System	Ac 6	Ac 6-1k	Ac 4	Ac 4-1k
LIF	0.672_{1}	0.516_{1}	0.725_{1}	0.692_{1}
ELiRF	0.659_{2}	0.488_{3}	0.722_{2}	0.645_{5}
GSI	0.618_{3}	0.487_{4}	0.690_{4}	0.658_{3}
DLSI	0.595_{4}	0.385_{14}	0.655_{6}	0.637_{7}
GTI-GRAD	0.592_{5}	0.509_{2}	0.695_{3}	0.674_{2}
LYS-LG•	$\mathbf{0.568}_{6}$	$\mathbf{0.434}_{5}$	$\mathbf{0.664}_{5}$	$\mathbf{0.634}_{9}$
DT	0.557_{7}	0.408_{10}	0.625_{7}	0.601_{11}
ITAINNOVA	0.549_{8}	0.405_{11}	0.610_{10}	0.484_{14}
BittenPotato	0.535_{9}	0.418_{8}	0.602_{11}	0.632_{10}
LYS-LSTM•	$\mathbf{0.505}_{9*}$	$\mathbf{0.430}_{6*}$	$\mathbf{0.599}_{11*}$	$\mathbf{0.605}_{10*}$
SINAI-ESMA	0.502_{10}	0.411_{9}	-	-
CU	0.495_{11}	0.419_{7}	0.481_{13}	0.600_{12}
INGEOTEC	0.488_{12}	0.431_{6}	-	-
SINAI	0.474_{13}	0.389_{13}	0.619_{8}	0.641_{6}
TID-SPARK	0.462_{14}	0.400_{12}	0.594_{12}	0.649_{4}
GAS-UCR	0.342_{15}	0.338_{15}	0.446_{14}	0.556_{13}
UCSP	0.273_{16}	-	0.613_{9}	0.636_{8}

Table 1: Comparison of accuracy for Task 1, between the best performance of each participant with respect to our machine- and deep learning models. Bold runs indicate our L2-LG and LSTM runs. Subscripts indicate the ranking for each group for their best run.

Finally, Table 4 compares the performance of the participating systems Task 2, both for

[1]The gradient signal becomes either too small or large causing a very slow learning or a diverging situation, respectively.

[2]http://deeplearning.net/tutorial/

[3]The size of the vector obtained for each word and the number of hidden units on the LSTM layer.

[4]Number of words to be indexed. The rest of the words are set to *unknown tokens*, giving to all of them the same index.

Corpus	N+	N	NEU	NONE	P	P+
L_6	0.000	0.486	0.000	0.582	0.049	0.575
L_6-1k	0.090	0.462	0.093	0.508	0.209	0.603
L_4	-	0.623	0.00	0.437	0.688	-
L_4-1k	-	0.587	0.00	0.515	0.679	-

Table 2: F1 score of our LSTM-RNN model for each test set proposed at Task 1. *1k* refers to the manually-labelled corpus containing 1 000 tweets.

Corpus	N+	N	NEU	NONE	P	P+
L_6	0.508	0.464	0.135	0.613	0.205	0.682
L_6-1k	0.451	0.370	0.000	0.446	0.232	0.628
L_4	-	0.674	0.071	0.569	0.747	-
L_4-1k	-	0.642	0.028	0.518	0.714	-

Table 3: F1 score of our L2-LG model for each test set proposed at Task 1

football and political tweets. The trend remains in this case and the machine learning approaches outperformed again our deep learning proposal.

6 Conclusions and future research

In the 4th edition of TASS 2015, we have tried a long short-term memory neural network to determine the polarity of tweets at the global and aspect levels. The performance of this model has been compared with the performance of our last-year system, based on an L2 logistic regression. Experimental results suggest that we need to explore new architectures and specific word embedding representations to obtain state-of-the-art results on sentiment analysis tasks. In this respect, we believe sentiment-specific word embeddings and other deep learning approaches (Tang et al., 2014) can help enrich our current model. Unsupervised pretraining has also been shown to improve performance of deep learning architectures (Severyn and Moschitti, 2015).

References

Bastien, F., P. Lamblin, R. Pascanu, J. Bergstra, I. Goodfellow, A. Bergeron, N. Bouchard, D. Warde-Farley, and Y. Bengio. 2012. Theano: new features and speed improvements. *arXiv preprint arXiv:1211.5590*.

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

Bespalov, D., B. Bai, Y. Qi, and A. Shokoufandeh. 2011. Sentiment classification based on supervised latent n-gram analysis. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 375—-382. ACM.

Brooke, J, M Tofiloski, and M Taboada. 2009. Cross-Linguistic Sentiment Analysis: From English to Spanish. In *Proceedings of the International Conference RANLP-2009*, pages 50–54, Borovets, Bulgaria. ACL.

Gimpel, K, N Schneider, B O'connor, D Das, D Mills, J Eisenstein, M Heilman, D Yogatama, J Flanigan, and N A Smith. 2011. Part-of-speech tagging for Twitter: annotation, features, and experiments. *HLT '11 Proc. of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers*, 2:42–47.

Hochreiter, S and J. Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Hurtado, L. and F. Pla. 2014. ELiRF-UPV en TASS 2014: Análisis de sentimientos, detección de tópicos y análisis de sentimientos de aspectos en twitter. In *Proceedings of the TASS workshop at SEPLN*.

Joshi, M and C Penstein-Rosé. 2009. Generalizing dependency features for opinion mining. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, ACLShort '09, pages 313–316, Stroudsburg, PA, USA. Association for Computational Linguistics.

System	SOCIAL-TV	STOMPOL
ELiRF	0.633_1	0.655_1
LYS-LG$^\bullet$	$\mathbf{0.599}_2$	$\mathbf{0.610}_4$
GSI	-	0.635_2
TID-SPARK	0.557_3	0.631_3
LYS-LSTM$^\bullet$	$\mathbf{0.540}_{3*}$	$\mathbf{0.522}_{4*}$

Table 4: Comparison of accuracy for Task 2, between the best run of the rest of participants and our machine and deep learning models

Mikolov, T., K. Chen, G. Corrado, and J. Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

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.

Montejo-Raéz, A., M. A. García-Cumbreras, and M. C. Díaz-Galiano. 2014. Participación de SINAI word2vec en TASS 2014. In *Proceedings of the TASS workshop at SEPLN*.

Pang, B. and L. Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 115–124. Association for Computational Linguistics.

Pang, B., L. Lee, and S Vaithyanathan. 2002. Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of EMNLP*, pages 79–86.

Pennebaker, J. W., M. E. Francis, and R. J. Booth. 2001. Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates*, page 71.

Saralegi, X. and I. San Vicente. 2013. Elhuyar at TASS 2013. In Alberto Díaz Esteban, Iñaki Alegría Loinaz, and Julio Villena Román, editors, *XXIX Congreso de la Sociedad Española de Procesamiento de Lenguaje Natural (SEPLN 2013). TASS 2013 - Workshop on Sentiment Analysis at SEPLN 2013*, pages 143–150, Madrid, Spain, September.

Severyn, A. and A. Moschitti. 2015. UNITN: Training Deep Convolutional Neural Network for Twitter Sentiment Classification. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 464–469, Denver, Colorado. Association for Computational Linguistics.

Socher, R., A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Ng, and C. Potts. 2013. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. In *EMNLP 2013. 2013 Conference on Empirical Methods in Natural Language Processing. Proceedings of the Conference*, pages 1631–1642, Seattle, Washington, USA. ACL.

Tang, D., F. Wei, N. Yang, M. Zhou, T. Liu, and B. Qin. 2014. Learning sentiment-specific word embedding for twitter sentiment classification. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 1555–1565.

Taulé, M., M. A. Martí, and M. Recasens. 2008. AnCora: Multilevel Annotated Corpora for Catalan and Spanish. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odjik, Stelios Piperidis, and Daniel Tapias, editors, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, pages 96–101, Marrakech, Morocco.

Thelwall, M., K. Buckley, G. Paltoglou, D. Cai, and A. Kappas. 2010. Sentiment Strength Detection in Short Informal Text. *Journal of the American Society for Information Science and Technology*, 61(12):2544–2558, December.

Vilares, D., M. A. Alonso, and C. Gómez-Rodríguez. 2015a. A linguistic approach for determining the topics of Spanish Twitter messages. *Journal of Information Science*, 41(2):127–145.

Vilares, D., M. A. Alonso, and C. Gómez-Rodríguez. 2015b. On the usefulness of lexical and syntactic processing in polarity classification of Twitter messages. *Journal of the Association for Information Science Science and Technology*, to appear.

Vilares, D., M. A Alonso, and C. Gómez-Rodríguez. 2015c. A syntactic approach for opinion mining on spanish reviews. *Natural Language Engineering*, 21(01):139–163.

Vilares, D., Y. Doval, M. A. Alonso, and C. Gómez-Rodríguez. 2014a. LyS at TASS 2014: A prototype for extracting and analysing aspects from spanish tweets. In *Proceedings of the TASS workshop at SEPLN*.

Vilares, D., M. Hermo, M. A. Alonso, C. Gómez-Rodríguez, and Y. Doval.

2014b. LyS : Porting a Twitter Sentiment Analysis Approach from Spanish to English na. In *Proceedings og The 8th InternationalWorkshop on Semantic Evaluation (SemEval 2014)*, number SemEval, pages 411–415.

Villena-Román, J., J. García-Morera, C. Moreno-García, S. Lana-Serrano, and J. C. González-Cristóba. TASS 2013 — a second step in reputation analysis in Spanish. *Procesamiento del Lenguaje Natural*, pages 37–44.

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

Villena-Román, L., E. Martínez-Cámara, Janine Morera-García, and S. M. Jiménez-Zafra. 2015. TASS 2014-the challenge of aspect-based sentiment analysis. *Procesamiento del Lenguaje Natural*, 54:61–68.

Zeiler, M.D. 2012. ADADELTA: An adaptive learning rate method. *arXiv preprint arXiv:1212.5701*.

TASS 2015, septiembre 2015, pp 53-58 recibido 20-07-15 revisado 24-07-15 aceptado 29-07-15

Ensemble algorithm with syntactical tree features to improve the opinion analysis

Algoritmo de ensamble con introducción de la estructura morfosintáctica para la mejora del análisis de opinión

Rafael del-Hoyo-Alonso
ITAINNOVA
C/ María de Luna, nº 7. 50018 Zaragoza, Spain
rdelhoyo@itainnova.es

María de la Vega Rodrigalvarez-Chamarro
ITAINNOVA
C/ María de Luna, nº 7. 50018 Zaragoza, Spain
vrodrigalvarez@itainnova.es

Jorge Vea-Murguía
ITAINNOVA
C/ María de Luna, nº 7. 50018 Zaragoza, Spain
jveamurguia@itainnova.es

Rosa María Montañes-Salas
ITAINNOVA
C/ María de Luna, nº 7. 50018 Zaragoza, Spain
rmontanes@itainnova.es

Resumen: En este artículo se describe cómo la integración en modo ensamble de varias técnicas de análisis de opinión permite mejorar la precisión del análisis u otros problemas de PLN donde el tamaño disponible del corpus de entrenamiento es pequeño en comparación al espacio de hipótesis y por tanto se deben explorar diversas estrategias. Dentro de las estrategias a introducir, se presenta una nueva forma de introducir el análisis morfo-sintáctico con objeto de encontrar relaciones que los métodos tradicionales no son capaces de realizar.
Palabras clave: Ensamble, árbol de Sintaxis, análisis de opinión. MaxEnt

Abstract: This article describes how the assemble of several opinion analysis techniques can improve the accuracy or in other NLP problems where the available size of the training corpus is small compared to the space of hypotheses and therefore should be explored a range of strategies. One of the strategies is based in a new way to include morpho-syntactic features to find relationships that traditional methods are not able to perform.
Keywords: Boosting, Sentiment Analysis, syntactic parser, MaxEnt

1 Introducción

El análisis de opinión es uno de los campos del análisis lingüístico donde las técnicas de minería de textos y de datos existentes es un excelente campo de batalla. Dentro del certamen TASS[1] 2015 (Villena-Román et al., 2015), se posibilita, a los diferentes grupos de investigación, compartir experiencias y permitir analizar el éxito de las diferentes aproximaciones. Este trabajo se enmarca dentro del reto primero de análisis de sentimiento.

En el análisis del sentimiento, la mayoría de las aproximaciones emplean algoritmos de aprendizaje automático (*Machine Learning*) (Pang et al., 2002) para construir clasificadores en base a un corpus predefinido donde se ha anotado la polaridad de la opinión manualmente. En esta dirección, la mayoría de los estudios (Feldman, 2013) buscan la manera mejor de obtener las características (*features*) del texto para introducir a un clasificador. Esta extracción de características de basan en diferentes transformaciones cómo pueden ser TF-IDF (del inglés Term frequency – Inverse document frequency), word2vector (Mikolov et al., 2013) o a través de la búsqueda de diccionarios donde se introduzca conocimiento previamente preestablecido por sicólogos o lingüistas y de de la búsqueda de expresiones polarizadas.

Encontrar un algoritmo de aprendizaje supervisado eficiente no es una tarea fácil, en especial cuando el problema es altamente complejo cómo en el caso del análisis del lenguaje natural. Cada algoritmo se caracteriza porque emplea una representación diferente de los datos. Encontrar una buena representación de estos algoritmos, que permita generalizar el problema a resolver, requiere de tiempo y de

[1] TASS (Taller de Análisis de Sentimientos en la SEPLN) website. http://www.daedalus.es/TASS.

múltiples ensayos previos. Cada método o aproximación al problema de análisis de opinión han demostrado que ciertas partes del espacio de datos es mejor modelarla por un método de clasificación en comparación a otros. El empleo de diferentes algoritmos o aproximaciones al problema puede proporcionar información complementaria importante sobre la representación de los datos, uno de los objetivos que queremos tratar con este artículo. Esto ha originado que utilizar una combinación o ensamble de clasificadores sea una buena alternativa para el análisis de un corpus lingüístico. Varios métodos para la creación de ensamble de clasificadores ya han sido propuestos, en análisis de opinión dentro del TASS 2013 (Martínez-Cámara et al., 2013).

Un ensamble de clasificadores se define cómo la utilización de un conjunto de clasificadores que combinan de alguna forma las decisiones individuales de cada uno de ellos para clasificar nuevas instancias (Dzeroski & Zenki, 2000). Esta técnica es común en el mundo del *Machine Learning*. Existen varias razones que justifican el ensamble de clasificadores. Algunas de éstas, son: i) los datos para el entrenamiento pueden no proveer suficiente información para elegir un único mejor clasificador debido a que el tamaño disponible en estos datos es pequeño en comparación al espacio de hipótesis, este es el caso más común que nos vamos a encontrar en el análisis del lenguaje natural y en especial en el caso del TASS; ii) la combinación redundante y complementaria de clasificadores mejora la robustez, exactitud y generalidad de toda la clasificación; iii) diferentes clasificadores utilizan diferentes técnicas y métodos de representación de los datos, lo que permite obtener resultados de clasificación con diferentes patrones de generalización; iv) los ensambles son frecuentemente mucho más exactos que los clasificadores individuales. Por el contrario, podemos encontrar que la solución global de diferentes clasificadores pueda enmascarar la existencia de un clasificador excelente. Existen diferentes técnicas de ensamble, *bagging, boosting, stacked generalization, random subspace method.* En la aproximación seguida hemos utilizado una técnica de *bagging*, entrenando cada algoritmo con una muestra del 80% del total, donde en vez de votar la mejor de las salidas se entrena un nuevo clasificador (*meta-learner*), al igual que en el modelo de *stacked generalization*. Cada algoritmo utilizado intenta explotar una variante del análisis de opinión con objeto de igualar el algoritmo de aprendizaje y de esta forma evitar desviaciones en el algoritmo de aprendizaje en el experimento, por el clasificador utilizado. El clasificador utilizado en todos los casos ha sido el modelo de clasificación de máxima entropía (*MaxEnt*). *MaxEnt* pertenece a la familia de los clasificadores exponenciales o *log-lineales* que extraen un conjunto de características (*features*) de la entrada, las combinan de modo lineal y utiliza el resultado cómo un exponente. *MaxEnt* es uno de los algoritmos más usados actualmente para análisis de Opinión.

Un paso más allá de la identificación de nuevas características o la introducción de intensificadores o negaciones en el análisis sintaxis del lenguaje es la introducción del análisis morfológico y análisis del árbol sintáctico dentro del análisis del sentimiento. Uno de los modelos qué más éxito ha tenido en el análisis y la captura de los efectos del árbol sintáctico es el modelo *Recursive Neural Tensor Network* (RNTN), (Socher et al., 2013) o modelos cómo *convolutional neural network* (Kalchbrenner et al., 2014), dentro de la familia de algoritmos denominados *deep learning*. RNTN captura el cambio en la sintaxis cuando se realiza un cambio en la opinión permitiendo alcanzar mayores tasas de éxito que los métodos tradicionales de n-gramas a cambio de necesitar un corpus de entrenamiento de gran tamaño y basado en frases cómo el *Stanford Sentiment Treebank* (Socher et al., 2013), no siendo por tanto extrapolable a la competición TASS. Inspirado en esa idea y con la introducción del ensamblado de algoritmos se ha paliado el efecto negativo del reducido tamaño del corpus de entrenamiento.

Finalmente diferentes pruebas y el prototipado de los algoritmos se ha realizado mediante la herramienta *Moriarty Analytics* del Instituto Tecnológico de Aragón.

En el siguiente artículo se describirá la arquitectura del algoritmo propuesto, explicando cada una de las aproximaciones utilizadas. A continuación, se realizará un resumen de los resultados obtenidos. Finalmente se describirán unas conclusiones y trabajo futuro.

2 Arquitectura del Algoritmo

La composición de algoritmos permite explorar aproximaciones distintas dentro del corpus de entrenamiento. Los algoritmos utilizados han sido 4 que intentan cubrir diferentes aspectos, cada uno de ellos se ha entrenado con un 80% del corpus distinto y un 20% se ha dejado para validar el resultado.

El primer algoritmo consiste en la utilización de una aproximación simple de *Machine Learning* tras una normalización básica del texto. El segundo algoritmo consiste en la utilización del algoritmo basado en diccionario Afectivo de Cynthia M. Whissell (Whissell et al., 1986). El tercer algoritmo consiste en la utilización de otro diccionario pero en este caso se utilizará un mayor número de dimensiones de la psicóloga Margaret M. Bradley (Bradley & Lang, 1999). Finalmente, como cuarto algoritmo, se introduce un último método que incorpora la información del árbol de sintaxis y la información morfológica.

Para ello todos los algoritmos poseen una etapa de normalización, esta normalización consiste en la lematización, la eliminación de *stopwords* (sin eliminar las negaciones), eliminación de URLs y menciones.

Con objetivo de evitar una desviación debida al clasificador se ha optado en todos los casos por utilizar el mismo y evitar las desviaciones motivadas por el tipo de clasificador. Este seleccionado ha sido el de máxima entropía (*MaxEnt*).

Finalmente la selección del resultado final fue realizada con un segundo clasificador que fue entrenado con el global de los datos. En este caso se seleccionó un algoritmo *Random Forest*, con objeto de poder interpretar los resultados obtenidos, evitar sobre ajuste al no ser un algoritmo paramétrico y por haber obtenido los mejores resultados con otros algoritmos en un 10 *crosss validation*.

2.1 Aproximación básica de Machine Learning

El primer algoritmo consiste en la introducción de un modelo de extracción de *features* automático mediante la utiliza del algoritmo tf-idf, y un máximo de 3 n-gramas. Estas *features* son introducidas al modelo de máxima entropía.

2.2 Diccionario Afectivo DAL

Esta aproximación se basa en la utilización de un diccionario afectivo, *Dictionary of Affect in Language* construido por Cynthia M. Whissell.

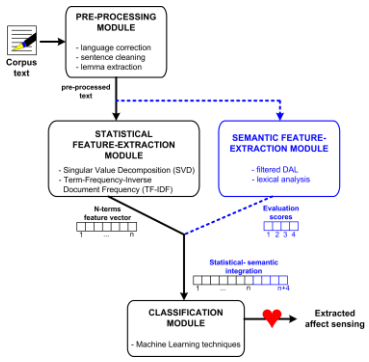

Figura 1 Arquitectura básica del algoritmo basado en DAL.

Una descripción del algoritmo utilizado puede verse en (Del-Hoyo et al., 2009).

2.3 Diccionario ANEW (Affective Norms for English Words)

Esta aproximación se basa en la utilización de un diccionario afectivo *Affective Norms for English Words* (ANEW; Bradley & Lang, 1999) y en este caso el diccionario afectivo en español, (Redondo et al., 2007). La utilización de este lexicón se ha utilizado por varios autores cómo (Gökçay et al., 2012; Gilbert, 2014)

El lexicón ANEW provee de un conjunto de 1.034 palabras etiquetadas emocionalmente. El diccionario de ANEW se ha etiquetado en tres dimensiones placer, activación y dominancia (pleasure, arousal, y dominance). La valencia va desde un valor de 1 hasta nuevo siendo 5 el valor considerado como neutro.

Existen diferentes formas de introducirlo por diferentes autores desde el análisis de la suma de la valencia del texto o introducir un clasificador de los textos en función de las 3 dimensiones de cada texto. En nuestro caso, discretizaremos cada una de las dimensiones en 5 etiquetas, y se definirán al igual que en el caso anterior de un nuevo número de características con las que entrenaremos el algoritmo de clasificación.

2.4 Introduciendo la estructura topológica en la ecuación

Otra forma para poder evaluar y realizar un análisis de opinión dentro de un texto es estudiando su estructura morfo-sintáctica.

En primer lugar, el texto ha sido dividido en frases. Utilizando las librerías proporcionadas por *Freeling* (Padró & Stanilovsky, 2012), se ha analizado cada una de las frases morfológicamente, identificando cada una de las palabras que forman parte de ella y etiquetándolas. Para el idioma castellano, las etiquetas utilizadas son las propuestas por el grupo *EAGLES*[2]. Por ejemplo, en la frase "*Hay varios cruceros atracados en el puerto de Santa Cruz*", el análisis morfológico queda de la siguiente forma:

Palabra	Etiqueta
Hay	VMIP3S0
varios	DI0MP0
cruceros	NCMP000
atracados	VMP00PM
en	SPS00
el	DA0MS0
puerto	NCMS000
de	SPS00
Santa Cruz	NP00G00

Tabla 1: Análisis Morfológico.

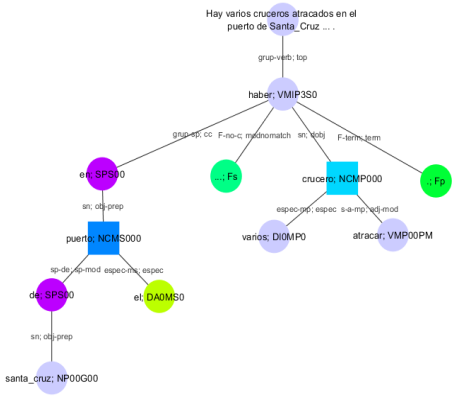

Figura 2: Árbol sintáctico obtenido.

Además, se genera el árbol sintáctico de cada una de las sentencias, generando un conjunto de grafos. El grafo resultante es el que se muestra en la figura 2.:

Una vez obtenido el grafo, donde se pueden ver de forma más sencilla las relaciones existente entre las diferentes palabras de una frase, se obtienen todos los caminos (o *paths*) existentes con cierto grado de profundidad. Siendo cada *path* una nueva característica a introducir al algoritmos *MaxEnt*.

Para detectar todos los *paths* existentes en una frase, se ha generado un algoritmo que detecta todos los nodos terminales del grafo, es decir, que no tienen ningún hijo y obtiene el camino existente entre el nodo inicial y el nodo final. Un *path* viene descrito por la propiedad *chunk* y *parser* que define la arista (unión entre dos vértices) y el análisis morfológico simplificado de la palabra (vértice)

Por ejemplo, para la rama: "*haber;en;puerto;de;Santa_Cruz*", el *path* resultante sería:

grup-verb;top/VM/grup-sp;cc/S/sn;obj-prep/NC00/sp-de;sp-mod/S/sn;obj-prep/NPG0

Todos los posibles *paths* obtenidos para esa rama con una longitud mínima de *x* y una longitud máxima de *y*, siendo el nivel de expansión en las pruebas realizadas de x=3 e y=5, serían las mostradas en la tabla siguiente.

Paths Posibles Rama
grup-verb;top/VM/grup-sp;cc
grup-verb;top/VM/grup-sp;cc/S
grup-verb;top/VM/grup-sp;cc/S/sn;obj-prep
VM/grup-sp;cc/S
……
sp-de;sp-mod/S/sn;obj-prep
sp-de;sp-mod/S/sn;obj-prep/NPG0
S/sn;obj-prep/NPG0

Tabla 2: Ejemplo de cómo se mapea un grafo en un conjunto de términos (posibles paths) del modelo.

Este proceso se seguiría para todas las ramas y se obtendrían todos los *paths* existentes para una frase. Por tanto todos los *paths* se incorporan cómo nuevas *features* para que, de esta forma, el algoritmo *MaxEnt* pueda encontrar las relaciones internamente no solamente de los pesos de las palabras por su semántica sino que pueda introducir la relación sintáctica y morfológica de las oraciones. Esta transformación, a diferencia de algoritmos cómo el RNTN que se desea mantener la estructura topológica de la entrada a través de la utilización de tensores, pretende mapear

[2] http://www.ilc.cnr.it/EAGLES96/home.html

espacios de dimensiones variables en un espacio de dimensión menor.

3 Resultados

A continuación en la tabla 3 se detallan la precisión en los datos test y desviación estándar esperada de cada uno de los algoritmos de forma independiente mediante la utilización de una evaluación *10 cross validation*.

	Basic	**ANEW**	**DAL**	**Syntactic**
Accu.	44.8%	43.8%	46.9%	41.3%
Desv.	1.1%	1.7%	1.6%	1.4%

Tabla 3: Resultados estimados en test obtenidos por cada una de los algoritmos mediante un 10 cross validation.

Se puede observar cómo la aproximación DAL obtiene unos mejores resultados en cambio, el algoritmo de análisis sintáctico obtiene peores resultados. Esto es debido a la explosión de nuevas características que genera este sistema, y para explotar mejor estas características, se necesita de un corpus mayor, cómo puede ser el *Sentiment Treebank*.

5 levels, Full test corpus				
	Accu.	**P**	**R**	**F1**
Ensem.	0.535	0.421	0.445	0.433
DAL.	0.524	0.419	0.397	0.408
5 levels, 1k corpus				
	Accu.	**P**	**R**	**F1**
Ensem.	0.405	0.401	0.392	0.396
DAL.	0.384	0.400	0.373	0.386
3 levels, Full test corpus				
	Accu.	**P**	**R**	**F1**
Ensem.	0.610	0.498	0.497	0.497
DAL.	0.594	0.484	0.482	0.483

Tabla 4: Resultados obtenidos en la prueba con el modelo ensembled y el mejor de los 4 modelos por separado, en cada una de las evaluaciones.

La composición de algoritmos se ha realizado a través de un modelo de Random Forest obteniendo una precisión de clasificación con los datos totales de entrenamiento de un 93%. Esto es así ya que se han utilizado un 80% de cada uno de ellos para el entrenamiento y todos ellos han obtenido alrededor de un 90% de precisión con los datos de entrenamiento, por tanto era de esperar encontrar esos datos de precisión. Pero sí que se ha analizado si la inclusión de los datos sintácticos mejora o no el sistema, y se ha visto que la precisión del algoritmo mejoraba en un 2% con respecto a no introducir la entrada correspondiente al algoritmo sintáctico, y esta mejora era mayor que la desviación encontrada en los datos obtenidos por las n iteraciones realizada por el *cross-validation*.

Para validar globalmente el sistema correctamente utilizamos la validación de TASS, para ello vamos a comparar el mejor de los algoritmos por separado de forma independiente al algoritmo compuesto. Se puede tabla 4 ver cómo en todos los casos el modelo compuesto gana al modelo DAL de forma independiente. Analizando los resultados por clases el sistema puede verse que el sistema compuesto se comporta de la misma forma que los sistemas individuales, obteniendo mejores resultados en las etiquetas Positivas que las Negativas, y fallando más en diferenciar la etiqueta NEU de NONE. Utilizando datos de evaluación se puede analizar cómo en determinados casos el modelo Sintáctico encuentra relaciones que no son encontradas por los otros modelos. Vamos a poner varios ejemplos a continuación.

"Y además no llueve! http://t.co/hXo9c8p9*"*

"1.Primer acto. Mario Monti entra en escena cambiando el panorama. Merkel con Berlusconi no hubiera confiado en las promesas de IT"

En ambos casos se obtiene el siguiente resultado:

	Basic	**ANEW**	**DAL**	**Sintactic**
Predic.	None	None	None	N

Tabla 5: Resultados de opinión negativa dónde la información morfo-sintáctica ofrece un valor añadido.

Puede verse que en frases donde existe una negación pero no existe ninguna palabra "afectiva" puede verse que subjetivamente es negativa. En cambio los otros algoritmos al no encontrar ningún vocabulario específico nos indican un valor nulo. Por el contrario, en varios casos el "Si" en oraciones condicionales enfatiza y modifica las predicciones.

"Jejeje. Si lo escuche. RT @cris_rc: "@SilviaSanz7: Jajajajaja Suena @AlejandroSanz en el (cont) http://t.co/OVHwMetV"

	Basic	**ANEW**	**DAL**	**Sintactic**
Predic.	P	P	P	P+

Tabla 6: Resultados de opinión dónde la información morfo-sintáctica ofrece un valor añadido.

4 Conclusiones

Se ha presentado la composición de algoritmos cómo herramienta para la mejora de la precisión del análisis de sentimiento u otros problemas de NLP donde el tamaño disponible del corpus de entrenamiento es pequeño en comparación al espacio de hipótesis y por tanto se deben explorar diversas estrategias. Por otra parte, se presentado una nueva forma de introducir el análisis morfo-sintáctico con objeto de encontrar nuevas relaciones.

Bibliografía

Bradley, M. M., & P. J. Lang, (1999). Affective norms for English words (ANEW): Instruction manual and affective ratings (pp. 1-45). Technical Report C-1, The Center for Research in Psychophysiology, University of Florida.

Del-Hoyo, R., I. Hupont, F. J. Lacueva, & D. Abadía, (2009, November). Hybrid text affect sensing system for emotional language analysis. In Proceedings of the international workshop on affective-aware virtual agents and social robots (p. 3). ACM.

Dzeroski, S. & B. Zenki (2000). Is Combining, Classifiers Better than Selecting the Best One.International Conference on Machine Learning (ICML): 123-130.

Feldman, R. (2013). Techniques and applications for sentiment analysis. Communications of the ACM, 56(4), 82-89.

Gilbert, C. H. E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text.

Gökçay, D., E. Işbilir, & G. Yıldırım, (2012, December). Predicting the sentiment in sentences based on words: An exploratory study on ANEW and ANET. In Cognitive Infocommunications, 2012 IEEE 3rd International Conference on (pp. 715-718)..

Kalchbrenner, N., E. Grefenstette, & P. Blunsom, (2014). A convolutional neural network for modelling sentences. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. June, 2014

Martinez-Cámara, E., Y. Gutiérrez-Vázquez, J. Fernández, A. Montejo-Ráez & R. Munoz-Guillena. (2013) Ensemble classifier for Twitter Sentiment Analysis. TASS 2013

Mikolov, T., K. Chen, G. Corrado, & J. Dean, (2013). Efficient estimation of word representations in vector space. in International Conference on Learning Representations

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

Pang, B., L. Lee, & S. Vaithyanathan, (2002, July). Thumbs up?: sentiment classification using machine learning techniques. In Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10 (pp. 79-86). Association for Computational Linguistics.

Redondo, J., I. Fraga, I. Padrón, & M. Comesaña, (2007). The Spanish adaptation of ANEW (affective norms for English words). Behavior research methods, 39(3), 600-605.

Socher, R., A. Perelygin, J. Y. Wu, J. Chuang, C. D. Manning, A. Y. Ng, & C. Potts, (2013, October). Recursive deep models for semantic compositionality over a sentiment treebank. In Proceedings of on empirical methods in natural language processing (EMNLP) (Vol. 1631, p. 1642).

Villena-Román, J., J. García-Morera, M. A.; García-Cumbreras, E. Martínez-Cámara, M. T. Martín-Valdivia, L. A. Ureña-López, (2015): Overview of TASS 2015, In Proceedings of TASS 2015: Workshop on Sentiment Analysis at SEPLN. vol. 1397.

Whissell, C., M. Fournier, R. Pelland, D. Weir, & K. A. Makarec, (1986) dictionary of affect in language. IV. Reliability, validity, and applications. Perceptual and Motor Skills, 62(3), 875-888.

TASS 2015, septiembre 2015, pp 59-64 recibido 20-07-15 revisado 24-07-15 aceptado 29-07-15

Participación de SINAI DW2Vec en TASS 2015*

SINAI DW2Vec participation in TASS 2015

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

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

Resumen: Este artículo describe el sistema de clasificación de la polaridad utilizado por el equipo SINAI-DW2Vec en la tarea 1 del workshop TASS 2015. Nuestro sistema no sigue el modelo espacio vectorial clásico ni aplica análisis sintáctico o léxico alguno. Nuestra solución se basa en un método supervisado con SVM sobre vectores de pesos concatenados. Dichos vectores se calculan utilizando dos técnicas: Word2Vec y Doc2Vec. La primera obtiene la sumatoria de vectores de palabras con un modelo generado a partir de la Wikipedia en español. Con la técnica Doc2Vec se generan vectores de características a partir de la colección de tweets de entrenamiento, en este caso a nivel de párrafo (o tweet) en lugar de a nivel de palabra como lo hace Word2Vec. La experimentación realizada demuestra que ambas técnicas conjuntas consiguen mejorar al uso de cada técnica por separado.
Palabras clave: Análisis de sentimientos, clasificación de la polaridad, deep-learning, Word2Vec, Doc2Vec

Abstract: This paper introduces the polarity classification system used by the SINAI-DW2Vec team for the task 1 at the TASS 2015 workshop. Our approach does not follow the Vector Space Model nor applies syntactic or lexical analyses. This solution is based on a supervised learning algorithm over vectors resulting from concatenating two different weighted vectors. Those vectors are computed using two different, yet related, algorithms: Word2Vec and Doc2Vec. The first algorithm is applied so as to generate a word vector from a deep neural net trained over Spanish Wikipedia. For Doc2Vec, the vector is generated with paragraph vectors (instead of word vectors) from a neural net trained over the tweets of the training collection. The experiments show that the combination of both vector distributions leads to better results rather than using them isolated.
Keywords: Sentiment analysis, polarity classification, deep learning, Word2Vec, Doc2Vec

1 Introducción

En este artículo describimos el sistema construido para participar en la tarea 1 del workshop TASS (Sentiment Analysis at global level), en su edición de 2015 (Villena-Román et al., 2015). Nuestra solución continúa con las técnicas aplicadas en el TASS 2014, utilizando aprendizaje profundo para representar el texto, y dando un paso más generando una representación no sólo a nivel de palabras sino también de frases o documentos. Para ello utilizamos el método *Word2Vec* utilizado con buenos resultados el año anterior, junto con la técnica *Doc2Vec* que nos permite representar un trozo variable de texto, por ejemplo una frase, en un espacio n-dimensional. Por lo tanto, utilizando Word2Vec generamos un vector para cada palabra del tweet, y realizamos la media de dichos vectores para obtener una única representación con Word2Vec. A dicho vector le concatenamos el vector obtenido con el modelo Doc2Vec, para generar una única representación del tweet. Una vez obtenidos los vectores de todos los tweets utilizamos un proceso de aprendizaje supervisado, a partir del conjunto de entrenamiento facilitado por la organización y el algoritmo SVM. Nuestros resultados demuestran que el uso conjunto de

* Esta investigación ha sido subvencionada parcialmente por el proyecto del gobierno español ATTOS (TIN2012-38536-C03-0), por la Comisión Europea bajo el Séptimo programa Marco (FP7 - 2007-2013) a través del proyecto FIRST (FP7-287607).

ambas técnicas mejora los resultados obtenidos utilizando sólo una de las técnicas presentadas.

Estos experimentos se presentan al amparo del TASS (Taller de Análisis de Sentimientos en la SEPLN), que es un evento satélite del congreso SEPLN, que nace en 2012 con la finalidad de potenciar dentro de la comunidad investigadora en tecnologías del lenguaje (TLH) la investigación del tratamiento de la información subjetiva en español. En 2015 se vuelven a proponer los mismos dos objetivos que en la convocatoria anterior. Por un lado observar la evolución de los sistemas de análisis de sentimientos, y por otro lado evaluar sistemas de detección de polaridad basados en aspectos.

La tarea del TASS en 2015 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 capítulo 2 describe el estado del arte de los sistemas de clasificación de polaridad en español. En el capítulo 3 se describe el sistema desarrollado y en el capítulo 4 los experimentos realizados, los resultados obtenidos y el análisis de los mismos. Finalmente, en el último capítulo 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 relevante 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). Otros sistemas, compararon diferentes técnicas de clasificación (Fernández Anta et al., 2012) implementadas en WEKA (Hall et al., 2009), o trataron la clasificación de forma binaria (Batista y Ribeiro, 2012), lanzando en paralelo distintos clasificadores binarios y combinando posteriormente los resultados. También se utilizó *naive-bayes multinomial* para construir un modelo del lenguaje (Trilla y Alías, 2012), un lexicón afectivo para representar el texto como un conjunto de emociones (Martín-Wanton y Carrillo de Albornoz, 2012), recuperación de información (RI) basado en divergencia del lenguaje para generar modelos de polaridad (Castellanos, Cigarrán, y García-Serrano, 2012), y un enfoque basado en el recurso léxico Sentitext, asignando una etiqueta de polaridad a cada término encontrado (Moreno-Ortiz y Pérez-Hernández, 2012).

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 alcanzaron 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. Otro equipo (Martínez Cámara et al., 2013) optó por una estrategia complemente no supervisada, frente a la supervisada desarrollada en 2012. Usaron como recursos lingüísticos SentiWordNet, Q-WordNet y iSOL, combinando los resultados y normalizando los valores.

3 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 cáculo del modelo Word2Vec hemos recurrido al software indicado, creado por los propios autores del método.

Basándose en Word2Vec, Le y Mikolov crearon el modelo Doc2Vec (Le y Mikolov, 2014). Este nuevo modelo calcula directamente un vector para cada párrafo o trozo de texto de longitud variable. El sistema para calcular dichos vectores es similar a Word2Vec, con la salvedad de que el contexto de cada palabra es inicializado en cada frase. Al igual que Word2Vec también existen dos topologías para dichos contextos de la palabras, bolsa de palabras distribuida (DCBOW) o memoria distribuida (DM - *Distributed Memory*).

Para calcular y utilizar el modelo Doc2Vec se ha utilizado una biblioteca para Python, denominada *gensim*[2]. Esta biblioteca también nos permite trabajar con el modelo Word2Vec generado anteriormente.

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[3] 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 200, logrando un modelo con más de 1,2 millones de palabras en su vocabulario.

Para crear el modelo de Doc2Vec hemos utilizado los propios tweets de entrenamiento y test. El motivo de esta decisión se debe principalmente a que la biblioteca Python para la creación de vectores Doc2Vec no nos ha permitido procesar toda la wikipedia (la misma que la utilizada para Word2Vec). Para utilizar los propios tweets hemos etiquetado cada uno con un identificador único que nos permita recuperar su vector del modelo. Además hemos generado un modelo con los siguientes parámetros: una ventana de 10 términos, el modelo DM y un número de dimensiones de 300. Estos parámetros se han eligido a partir de distintas pruebas empíricas realizadas con los tweets de entrenamiento.

Como puede verse en la Figura 1, nuestro sistema tiene tres 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. Otra en la que se entrena el modelo Doc2Vec con todos los tweets disponibles, tanto los tweets de entrenamiento como los de test. Y por último, otra en la que representamos cada tweet como la concatenación del vector obtenido con Doc2Vec y el vector como la media de los vectores Word2Vec de cada palabra en el tweet. Una simple normalización previa sobre el tweet es llevada a cabo, eliminando repetición de letras y poniendo todo a minúsculas. Así, el algoritmo SVM se entrena con un vector de 500 características como dimensión, resultado de dicha concatenación. La implementación de SVM utilizada es la basada en kernel lineal con entrenamiento SGD (Stochastic Gradient Descent) proporcionada por la biblioteca Sci-kit Learn[4] (Pedregosa et al., 2011).

Obtenemos así tres modelos: uno para los vectores de palabras según Wikipedia con

[1] `https://code.google.com/p/word2vec/`
[2] `http://radimrehurek.com/gensim/`
[3] `http://dumps.wikimedia.org/eswiki`
[4] `http://scikit-learn.org/`

Word2Vec, otro con los vectores de tweets según Doc2Vec, y otro para la clasificación de la polaridad con SVM. Esta solución es la utilizada en las dos variantes de la tarea 1 del TASS con predicción de 4 clases: la que utiliza el corpus de tweets completo (full test corpus) y el que utiliza el corpus balanceado (1k test corpus).

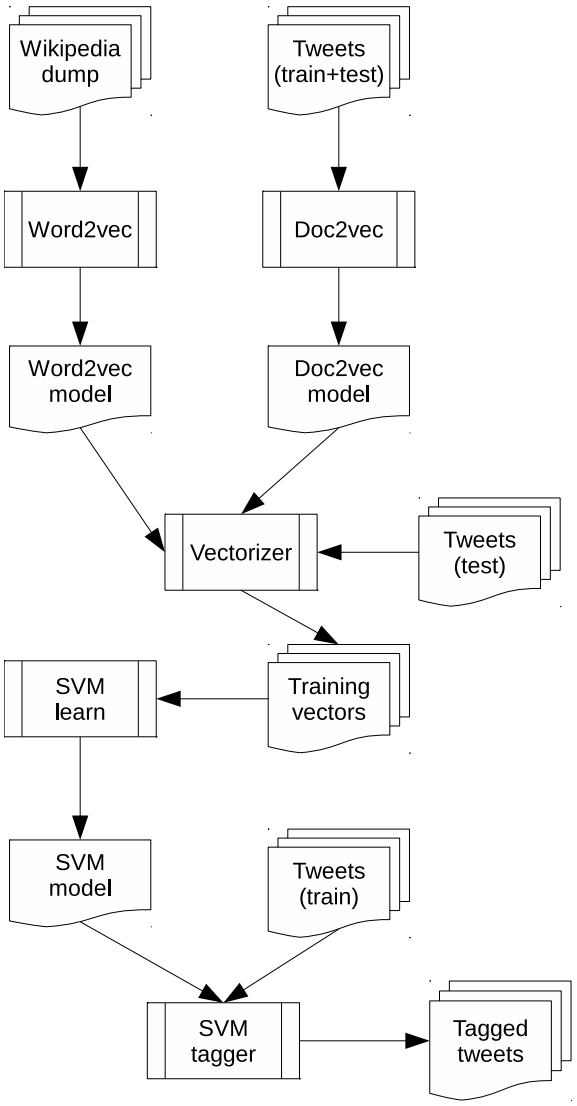

Figura 1: Flujo de datos del sistema completo

4 Resultados obtenidos

Para evaluar nuestro sistema hemos realizado diversos tipos de experimentos. Estos se diferencian según dos aspectos:

- Según el modelo utilizado para crear los vectores. Se han realizado experimentos utilizando sólo Word2Vec (*w2v*), sólo Doc2Vec (*d2v*) y concatenando los vectores de ambos modelos (*dw2v*).
- Según la colección de evaluación utilizada: Los organizadores pusieron a disposición de los participantes la colección completa (*full*) y una colección con un número de etiquetas más homogéneo que sólo contiene 1.000 tweets. Los experimentos con esta última colección han sido nombrados como *1k*.

Como se puede observar en la Tabla 1, los experimentos con mejores resultados son aquellos que utilizan los vectores generados por ambos modelos y la colección más homogenea llegando a alcanzar una precisión del **63 %** y un 46 % de Macro-F1. Con la colección completa, también se alcanzan los mejores resultados utilizando ambos modelos a la vez, obteniendo una precisión del **62 %** aproximadamente y un **47 %** de Macro-F1.

Modelo	Test coll	Accuracy	Macro-F1
wd2v	full	**0,619**	**0,477**
d2v	full	0,429	0,360
w2v	full	0,604	0,466
wd2v	1k	**0,633**	0,460
d2v	1k	0,510	0,306
w2v	1k	0,627	0,466

Tabla 1: Resultados obtenidos en los experimentos

Estos datos nos indican que, aún siendo un sistema bastante sencillo, se obtienen unos resultados prometedores. En ambas colecciones se han mejorado los resultados obtenidos con un único modelo (w2v y d2v) utilizando la conscatenación de ambos (wd2v). Sin embargo nuestra clasificiación no ha obtenido los resultados esperados, debido a que la mejora obtenida uniendo ambos modelos es muy pequeña en comparación con la utilización del modelo Word2Vec. Esto significa, que la utilización del modelos Doc2Vec en nuestros experimentos no es la correcta.

5 Conclusiones y trabajo futuro

Este trabajo describe una novedosa aplicación de los vectores de palabras generados por el método Word2Vec y Doc2Vec a la clasificación de la polaridad, consiguiendo una pequeña mejora en los resultados de precision y Macro-F1 en la competición TASS 2015, tarea 1. Estos resultados son destacables dada la simplicidad de nuestro sistema, que realiza un aprendizaje no supervisado para generar un modelo para representar cada tweet. No

obstante, existen diseños experimentales que no han podido ser acometidos y que esperamos poder realizar para evaluar mejor nuestro sistema, como por ejemplo utilizar una colección de tweets mucho mayor para entrenar el sistema Doc2Vec, o incluso la propia Wikipedia segmentada en frases o párrafos. Aunque para el uso de la Wikipedia con Doc2Vec es necesario un gran sistema computacional, nuestro primer objetivo sería reducir el número de párrafos seleccionando estos de forma aleatoria o utilizando alguna métrica de selección de características. De esta forma, podríamos observar si esta gran fuente de conocimiento es un recurso útil para Doc2Vec y posteriormente estudiar la manera de usar el recurso completo.

Los algoritmos de aprendizaje profundo prometen novedosas soluciones en el campo del procesamiento del lenguaje natural. Los resultados obtenidos con un modelo de palabras general no orientado a dominio específico alguno, ni a la tarea propia de clasificación de la polaridad, así como la no necesidad de aplicar técnicas avanzadas de análisis de texto (análisis léxico, sintáctico, resolución de anáfora, tratamiento de la negación, etc.) nos llevan a continuar nuestra investigación en una adecuación más específica de estos modelos neuronales en tareas concretas.

Es nuestra intención, por tanto, construir un modelo propio de aprendizaje profundo orientado a la clasificación de la polaridad. Gracias a los grandes volúmenes de datos, estas técnicas de aprendizaje profundo pueden aportar buenos resultados en este campo científico. En cualquier caso, es necesario un diseño cuidadoso de estas redes para lograr resultados mas ventajosos y cercanos a otros grupos que han participado en esta edición del TASS 2015, siendo este nuestro objetivo futuro.

Bibliografía

Batista, Fernando y Ricardo Ribeiro. 2012. The l2f strategy for sentiment analysis and topic classification. En *TASS 2012 Working Notes*.

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.

Castellanos, Angel, Juan Cigarrán, y Ana García-Serrano. 2012. Unedtass: Using information retrieval techniques for topic-based sentiment analysis through divergence models. En *TASS 2012 Working Notes*.

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.

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*.

Fernández Anta, Antonio, Philippe Morere, Luis Núñez Chiroque, y Agustín Santos. 2012. Techniques for sentiment analysis and topic detection of spanish tweets: Preliminary report. En *TASS 2012 Working Notes*.

Hall, Mark, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, y Ian H. Witten. 2009. The weka data mining software: An update. *SIGKDD Explor. Newsl.*, 11(1):10–18, Noviembre.

Le, Quoc V y Tomas Mikolov. 2014. Distributed representations of sentences and documents. *arXiv preprint arXiv:1405.4053*.

Martín-Wanton, Tamara y Jorge Carrillo de Albornoz. 2012. Uned en tass 2012: Sistema para la clasificación de la polaridad y seguimiento de temas. En *TASS 2012 Working Notes*.

Martínez Cámara, Eugenio, Miguel Ángel García Cumbreras, M. Teresa Martín Valdivia, y L. Alfonso Ureña López. 2013. Sinai-emml: Sinai-emml: Combinación de recursos lingüíticos para el análisis de la opinión en twitter. En *In Proc. of the TASS workshop at SEPLN 2013*.

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*.

Moreno-Ortiz, Antonio y Chantal Pérez-Hernández. 2012. Lexicon-based sentiment analysis of twitter messages in spanish. En *TASS 2012 Working Notes*.

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.

Trilla, Alexandre y Francesc Alías. 2012. Sentiment analysis of twitter messages based on multinomial naive bayes. En *TASS 2012 Working Notes*.

Villena-Román, Julio, Janine García-Morera, Miguel A. García-Cumbreras, Eugenio Martínez-Cámara, M. Teresa Martín-Valdivia, y L. Alfonso Ureña-López. 2015. Overview of tass 2015. En *In Proc. of TASS 2015: Workshop on Sentiment Analysis at SEPLN. CEUR-WS.org*, volumen 1397.

TASS 2015, septiembre 2015, pp 65-70 recibido 09-07-15 revisado 24-07-15 aceptado 29-07-15

Sentiment Analysis for Twitter: TASS 2015*

Análisis de sentimientos en Twitter: TASS 2015

Oscar S. Siordia
Daniela Moctezuma
CentroGEO
Col. Lomas de Padierna,
Delegación Tlalpan,
CP. 14240, México D.F.
osanchez;dmoctezuma@centrogeo.edu.mx

Mario Graff
Sabino Miranda-Jimenez
Eric S. Tellez
Elio-Atenógenes Villaseñor
INFOTEC
Ave. San Fernando 37,
Tlalpan, Toriello Guerra,
14050 Ciudad de México, D.F.
mario.graff;sabino.miranda;
eric.tellez;elio.villasenor@infotec.com.mx

Resumen: En este artículo se presentan los resultados obtenidos en la tarea 1: clasificación global de cinco niveles de polaridad para un conjunto de tweets en español, del reto TASS 2015. En nuestra metodología, la representación de los tweets estuvo basada en características lingüisticas y de polaridad como lematizado de palabras, filtros de palabras, reglas de negación, entre otros. Además, se utilizaron diferentes transformaciones como LDA, LSI y la matriz TF-IDF, todas estas representaciones se combinaron con el clasificador SVM. Los resultados muestran que LSI y la matriz TF-IDF mejoran el rendimiento del clasificador SVM utilizado.
Palabras clave: Análisis de sentimiento, Minería de opinión, Twitter

Abstract: In this paper we present experiments for global polarity classification task of Spanish tweets for TASS 2015 challenge. In our methodology, tweets representation is focused on linguistic and polarity features such as lemmatized words, filter of content words, rules of negation, among others. In addition, different transformations are used (LDA, LSI, and TF-IDF) and combined with a SVM classifier. The results show that LSI and TF-IDF representations improve the performance of the SVM classifier applied.
Keywords: Sentiment analysis, Opinion mining, Twitter.

1 Introduction

In last years the production of textual documents in social media has increased exponentially. This ever-growing amount of available information promotes the research and business activities around opinion mining and sentiment analysis areas. In social media, people share their opinions about events, other people and organizations. This is the main reason why text mining is becoming an important research topic. Automatic sentiment analysis in text is one of most important task in text mining. The task of sentiment classification determines if one document has positive, negative or neutral opinion or any level of each of them. Determining whether a text document has a positive or negative opinion is turning to an essential tool for both public and private companies (Peng, Zuo, y He, 2008). This tool is useful to know "What people think", which is an important information in order to help to any decision-making process (for any level of government, marketing, etc.) (Pang y Lee, 2008). With this purpose, in this paper we describe the methodology employed for the workshop TASS 2015 (Taller de Análisis de Sentimientos de la SEPLN). The TASS workshop is an event of SEPLN conference, which is a conference in Natural Language Processing for Spanish language. The purpose of TASS is to provide a discussion and a point of sharing about latest research work in the field

* This work was partially supported by Cátedras CONACYT program

of sentiment analysis in social media (specifically Twitter in Spanish language). In TASS workshop, several challenge tasks are proposed, and furthermore a benchmark dataset is proposed to compare the algorithms and systems of participants (for more details see (Villena-Román et al., 2015)).

Several methodologies to classify tweets from Task 1, Sentiment Analysis at global level of TASS workshop 2015, are presented in this work. This task is to perform an automatic sentiment classification to determine the global polarity (six polarity levels P, P+, NEU, N, N+ and NONE) of each tweet in the provided dataset. With this purpose, several solutions have been proposed in this work.

The paper is organized as follows, a brief overview of related works is shown in Section 2, the proposed methodology is describe in Section 3. Section 4 shows the experimental results and analysis, and finally, Section 5 concludes.

2 Related work

Nowadays, several methods have been proposed in the community of opinion mining and sentiment analysis. Most of these works employ twitter as a principal input of data and they aimed to classify entire documents as overall positive or negative polarity levels (sentiment) or rating scores (i.e. 1 to 5 stars).

Such is a case of work presented in (da Silva, Hruschka, y Jr., 2014), which proposes an approach to classify sentiment of tweets by using classifier ensembles and lexicons; where tweets are classified as positive or negative. As a result, this work concludes that classifier ensembles formed by several and diverse components are promising for tweet sentiment classification. Moreover, several state-of-the-art techniques were compared in four databases. The best accuracy result reported was around 75%.

In (Lluís F. Hurtado, 2014) is described the participation of ELiRF research group in TASS 2014 workshop (winners of TASS workshop 2014). Here, the winner approaches used for four tasks are detailed. The proposed methodology uses SVM (Support Vector Machines) with 1-vs-all approach. Moreover, Freeling (Padró y Stanilovsky, 2012) was used as lemmatizer and Tweetmotif[1] to tokenizer to Spanish language. The accuracy results of classification for task 1 are 64.32% (six labels) and 70.89% (four labels). F1 (F-Measure) is 70.48% in task 2 and 90% in task 3.

Another method to sentiment extraction and classification on unstructured text is proposed in (Shahbaz, Guergachi, y ur Rehman, 2014). Here, five labels were used to sentiment classification: Strongly Positive, Positive, Neutral, Negative and Strongly Negative. The solution proposed combines techniques of Natural language processing at sentence level and algorithms of opinion mining. The accuracy results were 61% for five levels and 75% by reducing to three levels (Positive, negative and neutral).

In (Antunes et al., 2011) an ensemble based on SVM and AIS (Artificial Immune Systems) is proposed. Here, the main idea is that SVM can be enhanced with AIS approaches which can capture dynamic models. Experiments were carried out with the Reuters-21578 benchmark dataset. The reported results show a 95.52% of F1.

An approach of multi-label sentiment classification is proposed in (Liu y Chen, 2015). This approach has three main components: text segmentation, feature extraction and multi-label classification. The features used included raw segmented words and sentiment features based on three sentiment dictionaries: DUTSD, NTUSD and HD. Moreover, here, a detailed study of several multi-label classification methods is conducted, in total 11 state-of-the-art methods have been considered: BR, CC, CLR, HOMER, RAkEL, ECC, MLkNN, and RF-PCT, BRkNN, BRkNN-a and BRkNN-b. These methods were compared in two microblog datasets and the reported results of all methods are around of 0.50 of F1.

In summary, most of works analyzed classify the documents mainly in three polarities: positive, neutral and negative. Moreover, most of works use social media (mainly Twitter) as analyzed documents. In this work, several methods to classify sentiment in tweets are described. These methods were implemented, according with TASS workshop specifications, with the purpose of classify tweets in six polarity levels: P+, P, Neutral, N+, N and None. The proposed method are based on several standard techniques as LDA (Latent Dirichlet Allocation), LSI (Latent Semantic Indexing), TF-IDF matrix in

[1] http://tweetmotif.com/about

combination with the well-known SVM classifier.

3 Proposed solution

In this section the proposed solution is detailed. First, a preprocessing step was carried out, later a Pseudo-phonetic transformation was done and finally the generation of Q-gram expansion was employed.

3.1 Preprocessing step

Preprocessing focuses on the task of finding a good representation for tweets. Since tweets are full of slang and misspellings, we normalize the text using procedures such as error correction, usage of special tags, part of speech (POS) tagging, and negation processing. Error correction consists on reducing words/tokens with invalid duplicate vowels and consonants to valid/standard Spanish words (ruidoooo → ruido; jajajaaa → ja; jijijji → ja). Error correction uses an approach based on a Spanish dictionary, statistical model for common double letters, and heuristic rules for common interjections. In the case of the usage of special tags, twitter's users (i.e., @user) and urls are removed using regular expressions; in addition, we classify 512 popular emoticons into four classes (P, N, NEU, NONE), which are replaced by a polarity tag in the text, e.g., positive emoticons such as `:)`, `:D` are replaced by `_POS`, and negative emoticons such as `:(`, `:S` are replaced by `_NEG`. In the POS-tagging step, all words are tagged and lemmatized using the Freeling tool for Spanish language (Padró y Stanilovsky, 2012), stop words are removed, and only content words (nouns, verbs, adjetives, adverbs), interjections, hashtags, and polarity tags are used for data representation. In negation step, Spanish negation markers are attached to the nearest content word, e.g., 'no seguir' is replaced by 'no_seguir', 'no es bueno' is replaced by 'no_bueno', 'sin comida' is replaced by 'no_comida'; we use a set of heuristic rules for negations. Finally, all diacritic and punctuation symbols are also removed.

3.2 Psudo-phonetic transformation

With the purpose of reducing typos and slangs we applied a semi-phonetic transformation. First, we applied the following transformations (with precedence from top to bottom):

$$
\begin{aligned}
cs|xc &\rightarrow x \\
qu &\rightarrow k \\
gue|ge &\rightarrow je \\
gui|gi &\rightarrow ji \\
sh|ch &\rightarrow x \\
ll &\rightarrow y \\
z &\rightarrow s \\
h &\rightarrow \epsilon \\
c[a|o|u] &\rightarrow k \\
c[e|i] &\rightarrow s \\
w &\rightarrow u \\
v &\rightarrow b \\
\Psi\Psi &\rightarrow \Psi \\
\Psi\Delta\Psi\Delta &\rightarrow \Psi\Delta
\end{aligned}
$$

In our transformation notation, square brackets do not consume symbols and Ψ, Δ means for any valid symbols. The idea is not to produce a pure phonetic transformation as in Soundex (Donald, 1999) like algorithms, but try to reduce the number of possible errors in the text. Notice that the last two transformation rules are partially covered by the statistical modeling used for correcting words (explained in preprocessing step). Nonetheless, this pseudo-phonetic transformation does not follow the statistical rules of the previous preprocessing step.

3.3 Q-gram expansion

Along with the placing bag of words representation (of the normalized text) we added the 4 and 5 gram of characters of the normalized text. Blank spaces were normalized and taken into account to the q-gram expansion; so, some q-grams will be over more than one word. In addition of these previous steps, several transformations (LSI, LDA and TF-IDF matrix) were conducted to generate several data models for testing phase.

4 Results and analysis

The classifier submitted to the competition was selected using the following procedure. The $7,218$ tweets with 6 polarity levels were split in two sets. Firstly, the tweets provided were shuffled and then the first set, hereafter the training set, was created with the first $6,496$ tweets (approximately 90% of dataset), and, the second set, hereafter the validation set, was composed by the rest 722

tweets (approximately 10% of dataset). The training set was used to fit a Support Vector Machine (SVM) using a linear kernel[2] with $C = 1$, weights inversely proportional to the class frequencies, and using one vs rest multiclass strategy. The validation set was used to select the best classifier using as performance the score F1.

The first step was to model the data using different transformations, namely Latent Dirichlet Allocation (LDA) using an online learning proposed by (Hoffman, Bach, y Blei, 2010), Latent Semantic Indexing (LSI), and TF-IDF.[3] Figure 1 presents the score F1, in the validation set, of a SVM using either LSI or LDA with normalized text, different levels of Q-gram (4-gram and 5-gram), and the number of topics is varied from 10 to 500 as well. It is observed that LSI outperformed LDA in all the configurations tested. Comparing the performance between normalized text, 4-gram, and 5-gram, it is observed an equivalent performance. Given that the implemented LSI depends on the order of the documents more experiments are needed to know whether any particular configuration is statistically better than other. Even though the best configuration is LSI with 400 topics and 5-gram, this system is not competitive enough compared with the performance presented by the best algorithm in TASS 2014.

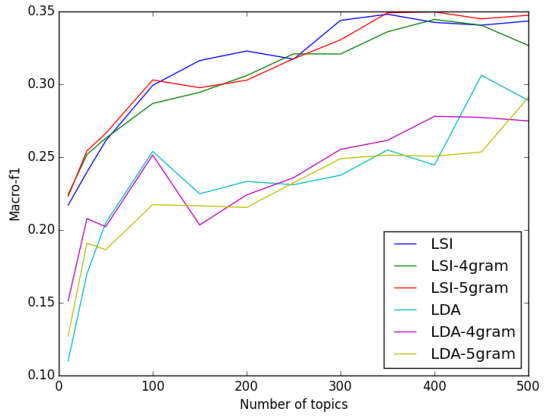

Figure 1: Performance in terms of the score F1 on the validation set for different number of topics using LSI and LDA with different Q-gram.

Table 1 complements the information presented on Figure 1.

The table presents the score F1 per polarity and the average (Macro-F1) for different configurations. The table is divided in five blocks, the first and second correspond to a SVM with LSI (400 topics) and TF-IDF, respectively. It is observed that TF-IDF outperformed LSI; within LSI and TF-IDF it can be seen that 5-gram and 4-gram got the best performance in LSI and TF-IDF, respectively.

The third row block presents the performance when the features are a direct addition of LSI and TF-IDF; here it is observed that the best performance is with 4-gram furthermore it had the best overall performance in N+. The forth row block complements the previous results by presenting the best performance of LSI and TF-IDF, that is, LSI with 5-gram and TF-IDF with 4-gram. It is observed that this configuration has the best overall performance in P+, N, None and average (Macro-F1). Finally, the last row block gives an indicated of whether the phonetic transformation is making any improvement. The conclusion is that the phonetic transformation is making a difference; however, more experiments are needed in order to know whether this difference is statistically significant.

Based on the score F1 presented on Table 1 the classifier submitted to the competition is a SVM with a direct addition of LSI using 400 topics and 4-gram and LDA with 5-gram. This classifier is identified as INGEOTEC-M1[4] in the competition. The SVM, LSI and LDA were trained with the 7218 tweets and then this instance was used to predict the 6 polarity levels of the competition tweets. This procedure was replicated for the 4 polarity levels competition.

Table 4 presents the accuracy, average recall, precision, and F1 of INGEOTEC-M1 run using the validation set created, a 10-fold crossvalidation on the 7218 tweets and the 1k tweets evaluated by the system's competition. This performance was on the 5 polarity levels challenge. It is observed from the table that the 10-fold crossvalidation gives a much better estimation of the performance

[2]The SVM was the class LinearSVC implemented in (Pedregosa et al., 2011)

[3]The implementations used for LDA, LSI, and TF-IDF were provided by (Řehůřek y Sojka, 2010).

[4]We also submitted another classifier identified as INGEOTEC-E1; however, the algorithm presented a bug that could not be find out on time for the competition.

	P	P+	N	N+	Neutral	None	Average
SVM + LSI							
Text	0.238	0.549	0.403	0.348	0.025	0.492	0.343
4-gram	0.246	0.543	0.404	0.333	0.048	0.533	0.351
5-gram	0.246	0.552	0.462	0.356	0.000	0.575	0.365
SVM + TF-IDF							
Text	0.271	0.574	0.414	0.407	0.103	0.511	0.380
4-gram	0.290	0.577	0.477	0.393	0.130	0.589	0.409
5-gram	0.302	0.577	0.476	0.379	0.040	0.586	0.393
SVM + {LSI + TF-IDF}							
4-gram	0.297	0.578	0.471	**0.421**	0.142	0.578	0.415
5-gram	0.307	0.567	0.474	0.391	0.040	0.579	0.393
SVM + {LSI with 4-gram + TF-IDF with 5-gram}							
4-5-gram	0.282	**0.596**	**0.481**	0.407	0.144	**0.595**	**0.417**
SVM + {LSI + TF-IDF without phonetic transformation}							
4-5-gram	**0.324**	0.577	0.459	0.395	**0.150**	0.593	0.416

Table 1: Score F1 per polarity level and average (Macro-F1) on the validation set for LSI (with 400 topics) and TF-IDF with different levels of Q-gram. The best performance in each polarity is indicated in boldface.

of the classifier when tested on 1k tweets of the competition (90% of training and 10% of validation).

In summary, in this work the best result reached was a 0.404 of F1. This result was achieved with a combination of LSI with 4-gram + TF-IDF with 5-gram, using a SVM classifier (one-vs-one approach).

	Acc.	Recall	Precision	F1
Val.	0.471	0.428	0.421	0.417
10-fold	0.443	0.397	0.395	0.393
Comp.	0.431	0.411	0.398	0.404

Table 2: Accuracy (Acc.), average recall, average precision and average F1 of the classifier in the validation set (Val.), using a 10-fold cross-validation ($7,218$ tweets), and as reported by the competition (comp.) on 1k tweets.

5 Conclusions

In this contribution, we presented the approach used to tackle the polarity classification task of Spanish tweets of TASS 2015. From the results, it is observed that a combination of different data models, in this case LSI and TF-IDF, improves the performance of a SVM classifier. It also noted that the phonetic transformation makes an improvement; however, more experiments are needed to know whether this improvement is statistically significant. As a result, we obtained a 0.404 of F1 (macro-F1) in sentiment classification task at five levels, with the proposed solution. This proposed solution uses a combination of LSI with 4-gram + TF-IDF with 5-gram, and a SVM classifier (one-vs-one approach).

Acknowledgements

This research is partially supported by the Cátedras CONACyT project. Furthermore, the authors would like to thank CONACyT for supporting this work through the project 247356 (PN2014).

Bibliography

Antunes, Mário, Catarina Silva, Bernardete Ribeiro, y Manuel Correia. 2011. A hybrid ais-svm ensemble approach for text classification. En *Adaptive and Natural Computing Algorithms*, volumen 6594 de *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, páginas 342–352.

da Silva, Nádia F.F., Eduardo R. Hruschka, y Estevam R. Hruschka Jr. 2014. Tweet sentiment analysis with classifier ensembles. *Decision Support Systems*, 66(0):170 – 179.

Donald, E Knuth. 1999. The art of computer programming. *Sorting and searching*, 3:426–458.

Hoffman, Matthew, Francis R Bach, y David M Blei. 2010. Online learning for latent dirichlet allocation. En *advances*

in neural information processing systems, páginas 856–864.

Liu, Shuhua Monica y Jiun-Hung Chen. 2015. A multi-label classification based approach for sentiment classification. *Expert Systems with Applications*, 42(3):1083 – 1093.

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

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.

Pang, Bo y Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135.

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.

Peng, Tao, Wanli Zuo, y Fengling He. 2008. Svm based adaptive learning method for text classification from positive and unlabeled documents. *Knowledge and Information Systems*, 16(3):281–301.

Řehůřek, Radim y Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. En *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, páginas 45–50, Valletta, Malta, Mayo. ELRA.

Shahbaz, M., A. Guergachi, y R.T. ur Rehman. 2014. Sentiment miner: A prototype for sentiment analysis of unstructured data and text. En *Electrical and Computer Engineering (CCECE), 2014 IEEE 27th Canadian Conference on*, páginas 1–7, May.

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

TASS 2015, septiembre 2015, pp 71-74 recibido 20-07-15 revisado 24-07-15 aceptado 29-07-15

BittenPotato: Tweet sentiment analysis by combining multiple classifiers.

BittenPotato: Análisis de sentimientos de tweets mediante la combinación de varios clasificadores.

Iosu Mendizabal Borda
(IIIA) Artificial Intelligence Research Institute
(CSIC) Spanish Council for Scientific Research
iosu@iiia.csic.es

Jeroni Carandell Saladich
(UPC) Universitat Politecnica de Catalunya
(URV) Universitat Rovira i Virgili
(UB) Universitat de Barcelona
jeroni.carandell@gmail.com

Resumen: En este artículo, usamos un saco de palabras (bag of words) sobre n-gramas para crear un diccionario de los atributos más usados en una dataset. Seguidamente, aplicamos cuatro distintos clasificadores, el resultado de los cuales, mediante diversas técnicas pretende mostrar la polaridad real de la frase extrayendo el sentimiento que contiene.
Palabras clave: Análisis de Sentimientos, Procesamiento de lenguaje natural.

Abstract: In this paper, we use a bag-of-words of n-grams to capture a dictionary containing the most used "words" which we will use as features. We then proceed to classify using four different classifiers and combine their results by apply a voting, a weighted voting and a classifier to obtain the real polarity of a phrase.
Keywords: Tweet sentiment analysis, natural language processing.

1 Introduction and objectives

Sentiment analysis is the branch of natural language processing which is used to determine the subjective polarity of a text. This has many applications ranging from the popularity of a certain product, the general opinion about an event or politician among many others.

In the particular case of twitter texts, these have the misfortune or great advantage of only consisting of a maximum of 140 characters. The disadvantage is that short texts aren't very accurately describable with bag of words which we will use, on the other hand, the limit also forces the author of the tweet to be concise in its opinion and therefore noise or non relevant statements are usually left out.

In this workshop for sentiment analysis focused on Spanish, a data set with tagged tweets according to their sentiment is given along with a description of evaluation measures as well as descriptions of the different tasks (Villena-Román et al., 2015).

The rest of the article is laid out as follows: Section 2 introduces the architecture and components of the system, namely the pre-processing, the extraction of features, the algorithms used and then the process applied to their results to obtain our final tag. Section 3 analyses the results obtained in this workshop. Finally, to conclude, in section 4 we will draw some conclusions and propose some future work.

2 Architecture and components of the system

Our system contains four main phases: data pre-processing, feature extraction - vectorization, the use of classifiers from which we extract a new set of features and finally a combined classifier which uses the latter to predict the polarity of the text.

2.1 Pre-processing

This step, crucial to all natural language processing task, consists of extracting noise from the text. Many of the steps such as removal of URLs, emails, punctuation, emoticons, spaced words etc. are general and we will not get into so much, yet some are more particular to the language in particular, such

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

as the removal of letters that are repeated more than twice in Spanish.

2.2 Vectorization: Bag of words

In order to be able to apply a classifier, we need to turn each tweet into a vector with the same features. To do this, one of the most common approach is to use the Bag-of-Words model with which given a corpus of documents, it finds the N most relevant words (or n-grams in our case). Each feature, therefore represents the appearance of a different relevant "word". Although the relevance of a word can be defined as the number of times it appears in the text, this has the disadvantage of considering words that appear largely throughout the whole document and lack semantic relevance. In order to counter this effect a more sophisticated approach called tf-idf (term frequency - inverse term frequency) is used. In our project we used the Scikit-Learn TfidfVectorizer (Pedregosa et al., 2011) to convert each tweet to a length N feature vector.

2.3 Algorithms: Classifiers

Once we have a way of converting sentences into a representation with the same features, we can use any algorithm for classification. Again, for all of the following algorithms we used the implementations in the Scikit-Learn python package (Pedregosa et al., 2011).

2.3.1 SVM

The first simple method we use is a support vector machine with a linear kernel in order to classify. This is generally the most promising in terms of all the used measures both with the complete of reduced number of classes.

2.3.2 AdaBoost (ADA)

Adaboost is also a simple, easy to train since the only relevant parameter is the number of rounds, and it has a strong theoretical basis in assuring that the training error will be reduced. However, this is only true with enough data (Freund and Schapire, 1999), given that the large number of features (5000) compared to the number of instances to train (around 4000 because of the cross validation with the training data that we will use to test the data), this is the worst performing method as can be seen in tables 1 and 2.

2.3.3 Random Forest (RF)

We decided to use this ensemble method as well because it has had very positive effects with accuracies that at times surpass the AdaBoosts thanks to its robustness to noise and outliers (Breiman, 2001).

2.3.4 Linear Regression (LR)

Since the degrees of sentiment polarity are ordered, we decided that it would also be appropriate to consider the problem as a discrete regression problem. Although a very straightforward approach, it seems to give the second best results in general at times surpassing the SVM (Tables 1 and 2).

2.4 Result: Combining classifiers

After computing the confusion matrices of the used classifiers we reached the conclusion that certain algorithms were better at capturing some classes than others. These confusion matrices can be observed in the next Section 3. Because of this reason we decided to combine the results of different classifiers to have more accurate results. In other words, we use the results of the single classifiers as a codification of the tweet into lower dimension. We can interpret each single classifier as an expert that gives its diagnose or opinion about the sentiment of a tweet. Since these different experts can be mistaken and disagree, we have to find the best result by combining the latter.

We tried three different combining methods. The first method is a simple voting of the different classifiers results and the more repeated one wins, in case of draws a random of the drawing ones will win. The second proposal is a more sophisticated voting with weights in each of the classifier results, these weights are computed with a train set and are normalized accuracies of the classification of this set.

Finally, the third method consists of another classification algorithm, this time of results. The idea is that we treat each previous classifier as an expert that give its own diagnose of the tweet, given that we have the real tweets, we decided to train a Radial Basis Function (RBF) with all of the training dataset and afterwards use the RBF to classify the final test results, which were the results we uploaded to the workshop. All three of these methods enhanced our results by few yet significant points. This can be thought of as a supervised technique for dimensionality

reduction, since we convert a dataset of 5000 features into only 4.

3 Empirical analysis

We are now going to analyse the results obtained in the workshop with the given testing tweet corpus. This section is separated in two subsections, firstly we will introduce the results obtained with the use of the four classifiers explained in Section 2.3. Secondly, we will focus on the usage of the three combining methods introduced in Section 2.4.

3.1 Single classifiers

First of all we are going to talk about the results obtained with the simple use of the four single classifiers explained in Section 2.3. The analysis is done with two different data sets; on the one hand a set separated in four classes and on the other hand a data set separated in six classes.

As it is depicted in Tables 1 and 2 the SVM and the Linear Regression classifiers are the most optimal ones in terms of the f1-measure which is the harmonic mean between the recall and the precision.

	Acc	Precision	Recall	F1
SVM	57.6667	0.4842	0.4759	0.4707
AB.	49.3333	0.4193	0.4142	0.4072
RF	54.0000	0.5122	0.4105	0.3968
LR	59.3333	0.4542	0.4667	0.4516

Table 1: Average measures in 3-Cross validated classifiers for 4 classes.

	Acc	Precision	Recall	F1
SVM	40.3333	0.3587	0.3634	0.3579
AB	35.0	0.3037	0.3070	0.2886
RF	39.3333	0.3370	0.3267	0.2886
LR	42.3333	0.3828	0.3621	0.3393

Table 2: Average measures in 3-Cross validated classifiers for 6 classes.

Observing the confusion matrix of the previously mentioned techniques, Random forest and Linear regression, we can learn perhaps more about the data itself. For instance, that the number of Neutral tweets are so low that tweets are rarely classified as such as seen in the NEU columns of the confusion matrices in figures 2 and 1. Another curious fact is that P+ labels are very separable for our classifiers. This could be because extremes might contain most key words that determine a positive review as opposed to the more subtle class P.

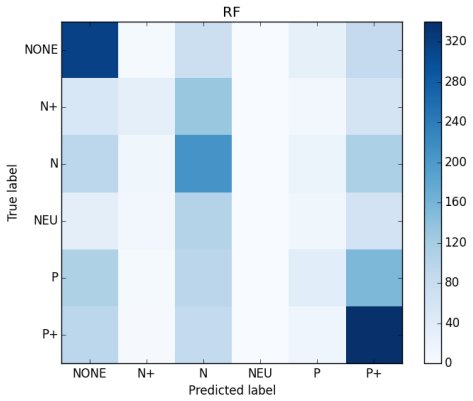

Figure 1: Confusion Matrix for a Random Forest with 6 classes.

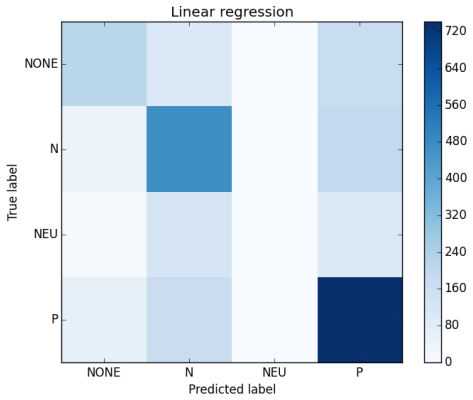

Figure 2: Confusion Matrix for Linear Regression with 4 classes.

3.2 Combining classifiers

After applying the 4 previous single classifiers to each tweet, we obtain a data matrix where each features correspond to the label set by each classifier. We can interpret this as some sort of dimensionality reduction technique where we now have a tweet transformed into an element of 4 attributes each corresponding to a classifier's results.

In tables 3 and 4 we can see the official results of the three combined classifiers on the Train data.

We have to keep in mind that when we are comparing the combined classifiers with the single classifiers, we are using two different

test sets. In the single classifiers, we use 3-Cross Validation exclusively on the train data to obtain average measure for each classifier. With the combined classifiers, we trained on the Train set and evaluated on the final Test set.

Notice that the weighted voting outperforms the normal voting. This seems intuitive because the weighted voting gives more importance to the most reliable classifiers. The RBF's results are not as promising as the previous two methods but it still outperforms all of the single classifiers.

	Acc	Precision	Recall	F1
Voting	59.3	0.500	0.469	0.484
Weighted Voting	59.3	0.508	0.465	0.486
RBF	60.2	0.474	0.471	0.472

Table 3: Official Results for the combined classifiers for 4 classes.

	Acc	Precision	Recall	F1
Voting	53.5	0.396	0.421	0.408
Weighted Voting	53.4	0.402	0.430	0.415
RBF	51.4	0.377	0.393	0.385

Table 4: Official results for the combined classifiers for 6 classes.

In general we can see that these methods, with the exception of the SVM in terms of F1-measure, outperform the rest.

4 Conclusions and future work

In this paper we have described our approach for the SEPLN 2015 for the global level with relatively good results considering the number of classes, and the general difficulty of the problem.

We have started by describing the initial preprocessing and the extraction of features using a bag of words on trigrams and bigrams. Then we have described and compared four different classifiers that we lated used as a way of translating the data into merely 4 dimensions, from 5000.

We can conclude that multiple classifiers are good at capturing different phenomena and that by combining them we tend to have a better global result as we have obtained in most of the TASS 2015 results of the Global level.

In general we are satisfied with the results obtained of the TASS2015 challenge. As future work, we propose to explore different classifiers that might capture different phenomena so that the combined classifier might have more diverse information. Also different combined classifiers should be trained.

References

Breiman, L. 2001. Random forests. *Machine Learning*, 45(1):5–32.

Freund, Y. and R. E. Schapire. 1999. A short introduction to boosting.

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.

Villena-Román, J., J. García-Morera, M. A. García-Cumbreras, E. Martínez-Cámara, M. T. Martín-Valdivia, and L. A. Ureña López. 2015. Overview of tass 2015.

TASS 2015, septiembre 2015, pp 75-79 recibido 20-07-15 revisado 24-07-15 aceptado 24-07-15

ELiRF-UPV en TASS 2015: Análisis de Sentimientos en Twitter

ELiRF-UPV at TASS 2015: Sentiment Analysis in Twitter

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

Davide Buscaldi
Université Paris 13
Sorbonne Paris Cité, LIPN
F-93430 Villetaneuse, France
davide.buscaldi@lipn.univ-paris13.fr

Resumen: En este trabajo se describe la participación del equipo del grupo de investigación ELiRF de la Universitat Politècnica de València en el Taller TASS2015. Este taller es un evento enmarcado dentro de la XXXI edición del Congreso Anual de la Sociedad Española para el Procesamiento del Lenguaje Natural. Este trabajo presenta las aproximaciones utilizadas para las todas las tareas del 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. Mediante esta técnica hemos conseguido mejorar las prestaciones de ediciones anteriores.
Palabras clave: Twitter, Análisis de Sentimientos.

Abstract: This paper describes the participation of the ELiRF research group of the Universitat Politècnica de València at TASS2015 Workshop. This workshop is a satellite event of the XXXI edition of the Annual Conference of the Spanish Society for Natural Language Processing. This work describes the approaches used for all the 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 this technique we have improved the performance of previous editions.
Keywords: Twitter, Sentiment Analysis.

1. Introducción

El Taller de Análisis de Sentimientos (TASS) ha venido planteando una serie de tareas relacionadas con el análisis de sentimientos en Twitter con el fin de comparar y evaluar las diferentes aproximaciones presentadas por los participantes. 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 cuarta edición del TASS se proponen dos tareas de ediciones anteriores (Villena-Román y García-Morera, 2013): 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 Social_TV, compuesto por tweets publicados durante la final de la Copa del Rey 2014. En esta edición del TASS 2015 (Villena-Román et al., 2015), se propone una tarea similar a la 2) pero utilizando un nuevo corpus llamado 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 2015, utilizan muchas de las características, desarrollos y recursos utilizados en las ediciones

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

en las que nuestro equipo ha participado (Pla y Hurtado, 2013) (Hurtado y Pla, 2014). El preproceso de los 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)[1]. También se ha usado *Freeling* (Padró y Stanilovsky, 2012)[2] 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*[3] y *LibLinear*[4]) que han demostrado ser eficientes implementaciones de SVM que igualan el estado del arte. El software se ha desarrollado en *Python* y para acceder a las librerías de SVM se ha utilizado el toolkit *scikit-learn*[5]. (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. Además, se ha incluido un nuevo recursos léxico, el diccionario *Afinn* (Hansen et al., 2011), que se ha traducido automáticamente del inglés al castellano y se ha adaptado para las tareas consideradas.

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.

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

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

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). Además se ha incorporado el diccionario *Afinn* traducido automáticamente del inglés al castellano y adaptado a la tarea.

Se han considerado tres alternativas para abordar la tarea:

- **run1** La primera alternativa combina mediante votación simple los 3 sistemas presentados en la edición del TASS de 2014.
- **run2-run4** La segunda alternativa explora diferentes combinaciones de parámetros y características de un modelo SVM. Para ello se han tenido en cuenta 192 configuraciones. A partir de éstas, se ha aprendido un segundo modelo SVM que sirve para proporcionar la nueva salida combinada. La diferencia entre el run2 y el run4 es que en el primero no se ajustan los parámetros del modelo SVM mientras que en el segundo, se elige una parte del entrenamiento para ajustar los parámetros. El sistema que se considera para la competición bajo esta aproximación es el run4.
- **run3** La tercera alternativa combina mediante un sistema de votación de mayoría simple las 192 configuraciones contempladas.

Para la subtarea de 4 etiquetas no se ha construido ningún sistema específico. Los tres sistemas enviados utilizan las salidas 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. Con los sistemas presentados se obtienen mejoras respecto a los resultados presentados en la edición anterior.

	Run	Accuracy
6-ETIQUETAS	run1	0.648
	run3	0.658
	run4	**0.673**
4-ETIQUETAS	run1	0.712
	run3	0.721
	run4	**0.725**

Tabla 2: Resultados oficiales del equipo *ELiRF-UPV* en la Tarea 1 de la competición TASS-2015 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 similar a la que se utilizó en nuestro sistema del TASS 2014, pero para esta edición hemos considerado 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.

La organización del TASS ha planteado dos subtareas. La primera utiliza el corpus *Social TV* y la segunda el corpus *STOMPOL*.

4.1. Corpus Social TV

El corpus Social_TV fue proporcionado por la organización y se compone de un conjunto de tweets recolectados durante la final de la Copa del Rey de fútbol de 2014. Está dividido en 1773 tweets de entrenamiento y 1000 tweets de test. El conjunto de entrenamiento está anotado con los aspectos y su correspondiente polaridad, utilizando en este caso sólo tres valores: P, N y NEU. El conjunto de test está anotado con los aspectos y se debe determinar la polaridad de éstos.

4.2. Corpus STOMPOL

El corpus STOMPOL se compone de un conjunto de tweets relacionados con una serie de aspectos políticos, como economía, sanidad, ...etc. que están enmarcado 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).

4.3. 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
SocialTV	run1	0.655
STOMPOL	run1	0.633

Tabla 3: Resultados oficiales del equipo *ELiRF-UPV* en la Tarea2 de la competición TASS-2015 para los corpus SocialTV y STOMPOL respectivamente.

En la tabla 3 se presentan los resultados obtenidos para la Tarea 2 sobre los dos corpus propuestos. Nuestra aproximación ha obtenido la primera posición en ambos corpus.

5. Conclusiones y trabajos futuros

En este trabajo se ha presentado la participación del equipo ELiRF-UPV en las 2 tareas planteadas en TASS 2015. Nuestro equipo ha utilizado técnicas de aprendizaje automático, en concreto, aproximaciones basadas en máquinas de soporte vectorial. Para ello hemos utilizado la librería para Python *scikit-learn* y las librerías externas *LibSVM* y *LibLinear*. Nuestra participación se ha centrado principalmente en explorar diferentes aproximaciones para combinar un conjunto de sistemas. Mediante esta técnica hemos conseguido mejorar las prestaciones de ediciones anteriores.

Nuestro grupo esta interesado en seguir trabajando en la minería de textos en redes sociales y especialmente en incorporar nuevos recursos a los sistemas desarrollados y estudiar nuevas estrategias y métodos de aprendizaje automático.

Como trabajo futuro nos planteamos desarrollar nuevos métodos de combinación de sistemas. También estamos interesados en considerar diferentes paradigmas de clasificación más heterogéneos (distintos de los SVM) para aumentar la complementariedad de los sistemas combinados.

Agradecimientos

Este trabajo ha sido parcialmente subvencionado por los proyectos DIANA: DIscourse ANAlysis for knowledge understanding (MEC TIN2012-38603-C02-01) y ASLP-MULAN: Audio, Speech and Language Processing for Multimedia Analytics (MEC 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.

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 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-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.

Villena-Román, Julio y Janine García-Morera. 2013. Workshop on sentiment analysis at sepln 2013: An over view. En *Proceedings of the TASS workshop at SEPLN 2013*. IV Congreso Español de Informática.

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

TASS 2015, septiembre 2015, pp 81-86 recibido 20-07-15 revisado 24-07-15 aceptado 29-07-15

Spanish Twitter Messages Polarized through the Lens of an English System

Mensajes de Twitter en español polarizados desde la perspectiva de un sistema para inglés

Marlies Santos Deas
Columbia University
New York, NY USA
ms5072@columbia.edu

Or Biran
Columbia University
New York, NY USA
orb@cs.columbia.edu

Kathleen McKeown
Columbia University
New York, NY USA
kathy@cs.columbia.edu

Sara Rosenthal
Columbia University
New York, NY USA
sara@cs.columbia.edu

Resumen: En este artículo describimos la adaptación al español de un sistema basado en aprendizaje automático con clasificación supervisada que fue desarrollado originalmente para el idioma inglés. El equipo de la Universidad de Columbia adaptó este sistema para participar en la Tarea 1 propuesta en TASS 2015, que consiste en determinar la polaridad a nivel global de un grupo de mensajes escritos en español en la red social Twitter.
Palabras clave: análisis de sentimiento, clasificación de polaridad

Abstract: In this paper we describe the adaptation of a supervised classification system that was originally developed to detect sentiment on Twitter texts written in English. The Columbia University team adapted this system to participate in Task 1 of the 4th edition of the experimental evaluation workshop for sentiment analysis focused on the Spanish language (TASS 2015). The task consists of determining the global polarity of a group of messages written in Spanish using the social media platform Twitter.
Keywords: sentiment analysis, polarity classification

1 Introduction

Sentiment analysis is the field concerned with analyzing the sentimental content of text. Most centrally, it involves the task of deciding whether an utterance contains *subjectivity*, as opposed to only objective statements, and determining the *polarity* of such subjective statements (e.g., whether the sentiment is positive or negative). Automatic sentiment analysis has important applications in advertizing, social media, finance and other fields. One variant that has become popular in recent years is sentiment analysis in microblogs, notably *Twitter*, which introduces difficulties common in that genre such as very short utterances, non-standard language and frequent out-of-vocabulary words.

The vast majority of work on sentiment analysis has been on English texts. Since methods for determining sentiment often rely on language-specific resources such as sentiment-tagged thesauri, they are often difficult to adapt to languages beyond English, as other language often have scarcer computational resources.

This paper describes the efforts of the Columbia University team at *Task 1* of TASS[1] 2015. TASS is an annual workshop focusing on sentiment analysis in Spanish, especially of short social media texts such as tweets. Each year, TASS proposes a number of tasks and collects the results of different participating systems.

In 2015, Task 1 is a combined subjectivity-polarity task: for each tweet, the competing system is expected to provide a label. There are two variants - the fine-grained variant, where there are six labels: {P+, P, Neu, N, N+, NONE}, and the coarse variant, where there are four labels: {P, Neu, N, NONE}. TASS distributes a standard data set of over 68,000 Spanish tweets for participants in this task (Villena-Román et al., 2015).

Instead of creating a new Spanish-specific system, we have adapted our existing English system to the Spanish language. We show that with relatively small engineering efforts and the proper resources, but without

[1] Taller de Análisis de Sentimientos

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

any language-specific feature engineering, our system can be adapted to a new language and achieve performance that is competitive with other systems at TASS. As a side effect, we formalized the process of adapting our system to any new language.

2 Related Work

Sentiment analysis in Twitter is a recent but popular task. In English, the SemEval Task of Sentiment Analysis in Twitter was the most popular task in both 2013 and 2014 (Rosenthal et al., 2014). In Spanish, TASS has organized a Twitter sentiment analysis task every year since 2012.

Multiple papers focusing on this task have been recently published. Most focus on supervised classification, using lexical and syntactic features (Go, Bhayani, and Huang, 2009; Pak and Paroubek, 2010; Bermingham and Smeaton, 2010). The latter, in particular, compare polarity detection in twitter to the same task in blogs, and find that despite the short and linguistically challenging nature of tweets, it is easier to detect polarity in Twitter than it is in blogs using lexical features, presumably because of more sentimental language in that medium.

Other work focused on more specialized features. Barbosa and Feng (2010) use a polarity dictionary that includes non-standard (slang) vocabulary words as well as Twitter-specific social media features. Agarwal et al. (2011) use the Dictionary of Affect in Language (DAL) (Whissell, 1989) and social media features such as slang and hashtags. Rosenthal, McKeown, and Agarwal (2014) use similar features, as well as features derived from Wiktionary, WordNet and emoticon dictionaries.

In Spanish, most work on Twitter sentiment analysis has been in the context of TASS. Many of the top-performing systems utilize a combination of lexical features, POS and specialized lexicons: the Elhuyar system relies on the Elhuyar Polar lexicon (Roncal and Urizar, 2014), while the LyS system (Vilares, Doval, and Gómez-Rodrıguez, 2014) and the CITIUS-CILENIS system (Gamallo and Garcia, 2013) each evaluate several Spanish-language lexicons. Other systems rely on distributional semantics (Montejo-Raez, Garcia-Cumbreras, and Diaz-Galiano, 2014) and on social media features (Zafra et al., 2014; Fernández et al., 2013).

3 Method

The main effort consisted of adapting an English sentiment analysis system for Spanish tweets, particularly for Task 1 of TASS 2015. The English system has been successfully applied to two editions of the SemEval Task 9 ("Sentiment Analysis in Twitter") - 2013 and 2014 (Rosenthal et al., 2014). The system consists of a Logistic Regression classifier that utilizes a variety of lexical, syntactic and specialized features (detailed in Section 3.2). It has two modes that can be run independently or in conjunction:

1. Subjectivity detection (distinguish between subjective and objective tweets)
2. Polarity detection (classify subjective tweets into positive, negative, or neutral).

The system is described in detail in Rosenthal and McKeown (2013). For the TASS task, four new modes were added:

1. Four-way classification, where the possible classes are P, N, NEU, and NONE
2. Four-way composite classification, where tweets are run through a two-step process: a binary classification (subjective, objective) followed by a three-way classification (P,N,NEU) of subjective tweets. Objective tweets in turn are given the label "NONE". Consequently, this two-step classification process yields to a four-way classifier. To train the subjectivity classifier, we grouped all labels other than "NONE" into one subjective label.
3. Six-way classification, where the possible classes are P, P+, N, N+, NEU, and NONE
4. Six-way composite classification (similar to four-way composite, and including two more labels: P+ and N+)

3.1 Preprocessing of tweets

Special tokens such as emoticons are replaced by a related word (e.g. "smiley") and supplemented with its affect values as represented in the DAL (Whissell, 1989). URLs and Twitter handles are converted to fixed tags that are not analyzed further to determine whether they are carriers of polarity. This process is unchanged from the English system.

We use the Stanford NLP library [2] for tag-

[2] http://nlp.stanford.edu/software/index.shtml

ging and parsing the tweet. Using the parse tree labels, we chunk the tweet into its shallow syntactic constituents (e.g. grup.nom). As in the English system, the chunker outputs one of three labels per token to indicate the position of the latter within a chunk: 'B' for beginning, 'I' for in (or intermediate, a continuation of the current chunk), and 'O' for out-of-vocabulary.

3.2 Features

The set of features used for Spanish is the same as that of the English system; we did not incorporate any Spanish-specific features for this task. The features currently utilized are essentially those described in Rosenthal, McKeown, and Agarwal (2014), and have evolved over time from the original system detailed in Agarwal, Biadsy, and Mckeown (2009).

In addition to lexical features (n-grams and POS), the system utilizes a variety of specialized features for social media text: emoticons; expanded web acronyms (LOL → laugh out loud) and contractions (xq → porque); punctuation and repeated punctuation; lengthened words in the tweet (e.g., largooooooo); all-caps words; and slang. We also use statistics of the DAL values for the words in the tweet (e.g., the mean activation, the max imagery, etc.).

3.3 Adaptation to Spanish

Adapting the English system to Spanish included two parts. First, we had to find Spanish equivalents to the English lexical resources (dictionaries, word lists etc.) that our system relies on. Second, we had to find equivalent Spanish NLP tools (a tokenizer, POS tagger and chunker).

3.3.1 Lexical Resources

The major challenge we faced was the lack of readily available resources in Spanish. In some cases, Spanish resources could be found and incorporated without a major effort - for example, the Spanish version of the DAL (Dell Amerlina Ríos and Gravano, 2013) was simple to integrate. In other cases, we had to put in more significant work - especially for the social media resources (e.g. the lists of contractions and emoticons). Table 1 details the English lexical resources used by our system and the Spanish equivalents, in addition to the location in which we found them or the method we used to create or adapt them.

We integrated the Standard Spanish dictionary distributed with Freeling[3] as our non-slang dictionary. For the DAL, we use the Spanish version created by Dell Amerlina Ríos and Gravano (2013). We leveraged the Google Translate service to create a Spanish version of our list of emoticons, and manually created a list of Spanish contractions.

The resulting Spanish resources are not identical to the original English ones. For example, the DAL scores for the word "abandon" and its Spanish translation "abandonar" are close but not exactly the same. Furthermore, the number of entries in the English DAL is more than three times that of the Spanish one, which results in a significant difference in coverage. In the standard dictionary, due to the highly inflected nature of the Spanish language, the number of entries more than quintuples when compared to the English version. Table 2 shows the percentage of the vocabulary (unique tokens) found in the training corpus for each resource. The standard dictionary has the highest coverage, followed by the DAL.

The English system utilizes a few additional resources, namely Wiktionary, WordNet and SentiWordNet. We have not yet integrated a Spanish version of these into the system, and consider that our first priority in future work. While Spanish equivalents of Wiktionary and WordNet do exist (Wikcionario and EuroWordNet), SentiWordNet does not have non-English counterparts. Our planned solution is to use MultiWordNet, a resource in which the English WordNet is aligned with other languages, to translate the English Synsets included in SentiWordNet into Spanish.

Resource	# Found	Percentage
Standard dictionary	10801	45.3 %
DAL	1845	7.7 %
NNP	8293	34.8 %
Punctuation and Numbers	259	1.1 %
Emoticons	11	∼ 0 %

Tabla 2: Coverage of the training set vocabulary by various resources

3.3.2 NLP Tools

We use the Spanish version of the Stanford Maxent Tagger (Toutanova and Manning,

[3]http://nlp.lsi.upc.edu/freeling

Resource	English	Spanish	Location or Method
Dictionary of Affect in Language (DAL)	abandon pleasantness mean: 1.0 activation mean: 2.38 imagery mean: 2.4 ...8,742 entries	abandonar 1.2 2.8 2.0 ...2,669 entries	http://habla.dc.uba.ar/gravano/sdal.php?lang=esp
Contractions	aren't → are not can't → cannot ...52 entries	pal → para el pallá → para allá ...21 entries	Manually created. Included variations with and without accent marks, apostrophes, and slang spelling.
Emoticons	:) happy :D laughter :-(sad ...99 entries	:) feliz :D risa :-(triste ...99 entries	Translated using Google translate
Standard dictionary (i.e. not slang)	98,569 entries, including proper nouns	556,647 entries, including inflected forms	Concatenated files contained in the Freeling installation and removed duplicates

Tabla 1: Parallel Lexical Resources

2000) for tagging. For chunking, we use the Spanish version of the Stanford Parser, and derive chunks from the lowermost syntactic constituents (or the POS, if the token is not a part of an immediate larger constituent).

For example, the Spanish phrase "Buen viernes" is chunked as follows:

Parse tree:
(ROOT (sentence (sn (grup.nom (s.a (grup.a (aq0000 Buen))) (w viernes)))))

Chunked phrase:
Buen/aq0000/B-grup.a viernes/w/B-w

4 Experiments and Results

We submitted two experiments (one simple and one composite; see Section 3) for each combination of classification task (four-way, six-way) and test corpus (full, 1k), for a total of eight experiments. The results are shown in Table 3. For each combination we show the accuracy and the macro-averaged precision, recall and F1 score.

We trained five models with the training data provided by TASS:

1. Four-way (P, N, Neu, NONE)
2. Six-way (P+, P, N+, N, Neu, NONE)
3. Subjectivity model (subjective, objective)
4. Three-way polarity (P, N, Neu)
5. Five-way polarity (P+, P, N+, N, Neu)

The last two were used in conjunction with the subjectivity model to form the composite classifier, as explained in Section 3.

4.1 Discussion

The task in which our system performs the best is the three-label classification using the joint four-way classifier described in Section 3. Both joint models (four-way and six-way) outperform their composite counterparts on the full test corpus. However, there is an improvement when using the composite model on the balanced 1k corpus, for both the three-label and five-label classification.

In terms of labels, our system consistently has the most difficulty classifying neutral (Neu) tweets across all experiments. In comparison, it did well in classifying strongly positive (P+) and objective (NONE) tweets, as well as positive (P) in the three-label subtask. Negative (N, N+) tweets were in between. Table 4 shows the performance of each system (for each task) on individual labels.

To assess the usefulness of our features in discriminating among the different classes, we looked at the odds ratios of the features for each class. Table 5 shows a few of the most discrimative features from each category: n-grams, POS and social media (SM). We found that social media features dominate across all classes, which is not a surprising outcome given the popular use of such features in Twitter communication. As shown in Table 5, emoticons such as a smiley face can be highly discriminative between positive and negative tweets, with a significantly stronger association with the former. Polar N-grams such as "felices" (happy) also constitute a relevant group and tend to be discriminative for the polar classes N and P. In the POS group, interrogative pronouns (pt) mar-

Task	TestSet	Variant	Acc.	Prec.	Rec.	F1	System Rank	Group Rank
5 Labels	Full	Six-way	**0.495**	0.393	0.441	**0.416**	28 / 37	11 / 16
		Composite	0.362	0.313	0.334	0.323	34 / 37	
	1k	Six-way	0.397	0.345	0.372	0.358	23 / 32	
		Composite	**0.419**	0.365	0.372	**0.369**	14 / 32	7 / 16
3 Labels	Full	Four-way	**0.597**	0.492	0.503	**0.497**	26 / 39	13 / 15
		Composite	0.481	0.404	0.403	0.404	36 / 39	
	1k	Four-way	0.578	0.450	0.493	0.470	31 / 39	
		Composite	**0.600**	0.461	0.481	**0.471**	25 / 39	13 / 16

Tabla 3: Sentiment Analysis results at global level (all measures are macro-averaged)

Task	Test Corpus	Variant	P	P+	N	N+	NEU	NONE
5 Labels	Full	Six-way	0.160	0.577	0.434	0.413	0.123	0.599
		Composite	0.096	0.490	0.370	0.278	0.088	0.376
	1k	Six-way	0.194	0.566	0.399	0.332	0.081	0.456
		Composite	0.260	0.583	0.438	0.335	0.078	0.493
3 Labels	Full	Four-way	0.676	N/A	0.603	N/A	0.108	0.544
		Composite	0.576	N/A	0.493	N/A	0.074	0.376
	1k	Four-way	0.667	N/A	0.584	N/A	0.079	0.483
		Composite	0.695	N/A	0.595	N/A	0.088	0.493

Tabla 4: F-measure of each class

king words such as "qué" (what) and "dónde" (where) are most important across all categories, followed by various types of verbs including semiauxiliary gerunds (vsg) and past indicative auxiliary (vais).

Group	P	N	NEU
Social Media	u	lol	:\(
	:D	u	\$
	\$	:D	k
n-grams	esfuerzo	esfuerzo	en @telediariointer 20:30
	gracias	petición de	))))
	felices	pide a rajoy	petición de
Part of Speech	pt000000	pt000000	pt000000
	vsg0000	vais000	vaif000
	vssp000	vsg0000	vsg0000

Tabla 5: Features with high Odds Ratios per class in four-way classification joint model

While it is difficult to compare our system's Spanish results with the results on English - the TASS dataset is quite different from the SemEval dataset - it is evident that the Spanish task is harder. This is not surprising, since we have fewer resources, and the ones which were adapted are in some cases not as comprehensive. However, the fact that we can get competitive results in Spanish using a system that was originally designed for English sentiment analysis shows that relatively quick and painless adaptation to other languages is possible.

5 Conclusion

We have adapted a sentiment analysis system, the original target language of which was English, to classifying the subjectivity and polarity of tweets written in Spanish for participation in Task 1 of TASS 2015. The English system provided significant leverage, allowing for direct reuse of most of its components, from the processing pipeline down to the features used by the classifier. The experimental results are encouraging, showing our system to be competitive with others submitted to TASS despite being adapted into Spanish from another language. From here on we will pursue further enhancements.

The main challenge we encountered was the need to substitute several English lexical resources that the system extensively employs with analogous Spanish variants that were not always easily attainable. In future work, we will incorporate the final missing pieces - Spanish versions of Wiktionary, WordNet and SentiWordNet - so that our Spanish system uses equivalents of all resources used by the English system.

While adapting our system to Spanish, we have compiled a list of necessary resources and presented some automated methods of quickly attaining such reasources in other languages (e.g., using Google Translate to quickly convert a list of emoticons). These along with resources and tools that we expect to be able to find for most languages (e.g.,

a standard dictionary and a list of contractions; a POS tagger and a constituent parser) comprise the bulk of the list. Some resources, such as the DAL, will potentially present a bigger challenge in other languages, but can possibly be automated through token translation as well. In future work, we will experiment with our system in additional languages and further refine our adaptation process.

Acknowledgements

This paper is based upon work supported by the DARPA DEFT Program. The views expressed are those of the authors and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

Bibliography

Agarwal, A., F. Biadsy, and K. R. McKeown. 2009. Contextual phrase-level polarity analysis using lexical affect scoring and syntactic n-grams. In *EACL*.

Agarwal, A., X. Boyi, I. Vovsha, O. Rambow, and R. Passonneau. 2011. Sentiment analysis of Twitter data. In *Workshop on Languages in Social Media*.

Barbosa, L. and J. Feng. 2010. Robust sentiment detection on Twitter from biased and noisy data. In *COLING*.

Bermingham, A. and A. F. Smeaton. 2010. Classifying sentiment in microblogs: Is brevity an advantage? In *International Conference on Information and Knowledge Management*.

Dell Amerlina Ríos, M. and A. Gravano. 2013. Spanish DAL: A Spanish dictionary of affect in language.

Fernández, J., Y. Gutiérrez, J. M. Gómez, P. Martinez-Barco, A. Montoyo, and R. Muñoz. 2013. Sentiment Analysis of Spanish Tweets Using a Ranking Algorithm and Skipgrams.

Gamallo, P. and M. García. 2013. TASS: A Naive-Bayes strategy for sentiment analysis on Spanish tweets. In *TASS 2013*.

Go, A., R. Bhayani, and L. Huang. 2009. Twitter sentiment classification using distant supervision.

Montejo-Raez, A., M.A. Garcia-Cumbreras, and M.C. Diaz-Galiano. 2014. Participación de Sinai Word2vec en TASS 2014. In *TASS 2014*.

Pak, E. and P. Paroubek. 2010. Twitter as a corpus for sentiment analysis and opinion mining. In *Conference on International Lang. Resources and Evaluation*.

Roncal, I. S. V. and X. S. Urizar. 2014. Looking for features for supervised tweet polarity classification. In *TASS 2014*.

Rosenthal, S. and K. McKeown. 2013. Columbia NLP: Sentiment detection of subjective phrases in social media.

Rosenthal, S., K. McKeown, and A. Agarwal. 2014. Columbia NLP: Sentiment detection of sentences and subjective phrases in social media. In *SemEval*.

Rosenthal, S., P. Nakov, A. Ritter, and V. Stoyanov. 2014. Semeval-2014 task 9: Sentiment analysis in Twitter.

Toutanova, K. and C. D. Manning. 2000. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *Joint SIGDAT Conference on Empirical Methods in NLP*.

Vilares, D., Y. Doval, and C. Gómez-Rodríguez. 2014. LyS at TASS 2014: A prototype for extracting and analysing aspects from Spanish tweets. In *TASS 2014*.

Villena-Román, J., J. García-Morera, M. A. García-Cumbreras, E. Martínez-Cámara, M. T. Martín-Valdivia, and L. A. Urena-Lopez. 2015. Overview of TASS 2015.

Whissell, C. 1989. The dictionary of affect in language. *Emotion: Theory, research, and experience*, 4.

Jiménez Zafra, S. M., E. Martínez Cámara, M.T. Martín Valdivia, and L.A. Ureña López. 2014. Sinai-esma: An unsupervised approach for sentiment analysis in Twitter. In *TASS 2014*.

TASS 2015, septiembre 2015, pp 87-92 recibido 20-07-15 revisado 24-07-15 aceptado 29-07-15

Comparing Supervised Learning Methods for Classifying Spanish Tweets

Comparación de Métodos de Aprendizaje Supervisado para la Clasificación de Tweets en Español

Jorge Valverde, Javier Tejada and Ernesto Cuadros
Universidad Católica San Pablo
Quinta Vivanco S/N , Urb. Campiña Paisajista, Arequipa - Perú
{andoni.valverde, jtejadac, ecuadros}@ucsp.edu.pe

Resumen: El presente paper presenta un conjunto de experimentos para abordar la tarea de clasificación global de polaridad de tweets en español del TASS 2015. En este trabajo se hace una comparación entre los principales algoritmos de clasificación supervisados para el Análisis de Sentimientos: Support Vector Machines, Naive Bayes, Entropía Máxima y Árboles de Decisión. Se propone también mejorar el rendimiento de estos clasificadores utilizando una técnica de reducción de clases y luego un algoritmo de votación llamado Naive Voting. Los resultados muestran que nuestra propuesta supera los otros métodos de aprendizaje de máquina propuestos en este trabajo.
Palabras clave: Análisis de Sentimientos, Métodos Supervisados, Tweets Españoles

Abstract: This paper presents a set of experiments to address the global polarity classification task of Spanish Tweets of TASS 2015. In this work, we compare the main supervised classification algorithms for Sentiment Analysis: Support Vector Machines, Naive Bayes, Maximum Entropy and Decision Trees. We propose to improve the performance of these classifiers using a class reduction technique and then a voting algorithm called Naive Voting. Results show that our proposal outperforms the other machine learning methods proposed in this work.
Keywords: Sentiment Analysis, Supervised Methods, Spanish Tweets

1 Introduction

Sentiment analysis is the computational study of opinions about entities, events, people, etc. Opinions are important because they often are taken into account in decision process. Currently, people use different social networks to express their experiences with products or commercial services. Twitter is one of the biggest repositories of opinions and it is also used as a communication channel between companies and customers. The data generated in Twitter is important for companies, because -- with that information -- they could know what is been saying about their products, services and competitors. In recent years, several researches of NLP have developed different methods to address the sentiment analysis problem in Twitter. The vast majority of works aim to classify a comment, according to the polarity expressed, in three categories: positive, negative or neutral (Koppel and Schler, 2006). The supervised classification algorithms are the most used methods to classify comments or opinions.

In this paper, we present a comparison of some supervised learning methods which have achieved good results in other research works. Analyzing the errors of those methods, we propose to use a class reduction technique and a voting algorithm (which take into account the results of supervised classifiers) to improve the classification of opinions in Twitter.

The rest of the paper is organized as follows: Section 2 summarizes the main works in

sentiment analysis. Section 3 describes our proposal and in Section 4 we describe the results that we have gotten. Finally, in Section 5, the conclusions of this work are presented.

2 Related Work

There are two general approaches to classify comments or opinions in positive, negative or neutral: supervised and unsupervised algorithms. Supervised classification algorithms are used in problems which are known a priori the number of classes and representative members of each class. The unsupervised classification algorithms, unlike supervised classification, do not have a training set, and they use clustering algorithms to try to create clusters or groups (Mohri, Rostamizadeh and Talwalkar, 2012).

The sentiment classification task could be formulated as a supervised learning problem with three classes: positive, negative and neutral. The most used supervised techniques in sentiment analysis are Naive Bayes (NB), Support Vector Machines (SVM), Maximum Entropy, etc. In most cases, SVM have shown great improvement over Naive Bayes.

Cui, Mittal, and Datar (2006) affirm that SVM are more appropriate for sentiment classification than generative models, because they can better differentiate mixed feelings. However, when the training data is small, a Naive Bayes classifier could be more appropriate. One of the earliest researches on supervised algorithms which classify opinions is presented in (Pang, Lee, and Vaithyanathan, 2002). In that work, authors use three machine learning techniques to classify the sentiment in movies comments. They test several features to find the most optimal set of them. Unigrams, bigrams, adjectives and position of words are used as features in those techniques. Ye, Zhang, and Law (2009) used three supervised learning algorithms to classify comments: SVM, Naive Bayes and Decision Trees. They use the frequencies of words to represent a document.

Most researches are focused for the English language, since it is the predominant language on the Internet. There are less works of sentiment analysis in Spanish opinions; however, Spanish is playing an important role. For Spanish comments, Perea-Ortega and Balahur (2014) present several experiments to address the global polarity classification task of Spanish tweets. Those experiments have focused on different feature replacements. The replacements were mainly based on repeated punctuation marks, emoticons and sentiment words. The proposal of Hernandez and Li (2014) is based on semantic approaches with linguistic rules for classifying polarity texts in Spanish. Montejo-Raez, Garcia-Cumbreras and Diaz-Galiano (2014) use supervised learning with SVM over the sum of word vectors in a model generated from the Spanish Wikipedia. Jimenez et al., (2014) developed an unsupervised classification system which uses an opinion lexicon and syntactic heuristic to identify the scope of Spanish negation words. San Vicente and Saralegi (2014) implement a Support Vector Machine (SVM) algorithm. That system combines the information extracted from polarity lexicons with linguistic features. For Peruvian Spanish opinions, Lopez, Tejada and Thelwall (2012) use a specialized dictionary with vocabulary of that country for Facebook comments. Lopez, Tejada and Thelwall (2012) proposed one of the first researches that analyze Peruvian opinions. In that work, authors use a basic method based on lexical resources to classify comments from Facebook.

3 Proposed Approach

This paper has two major objectives: First, we make a comparison of some of the main algorithms of supervised classification for Sentiment Analysis: Support Vector Machines, Naive Bayes, Maximum Entropy and Decision Trees. The second goal is to use a class reduction technique and then a voting algorithm to improve the accuracy of final results. The architecture of our system can be seen in Figure 1.

3.1 Comparison of Methods

In this paper we compare some classification methods in order to determine the performance of these algorithms in a set of opinions written by Spanish users. For the experiments, we used the four supervised classifiers described previously. The comparison of methods has the Training and Classification Phase. These phases will be explained below.

3.1.1 Training

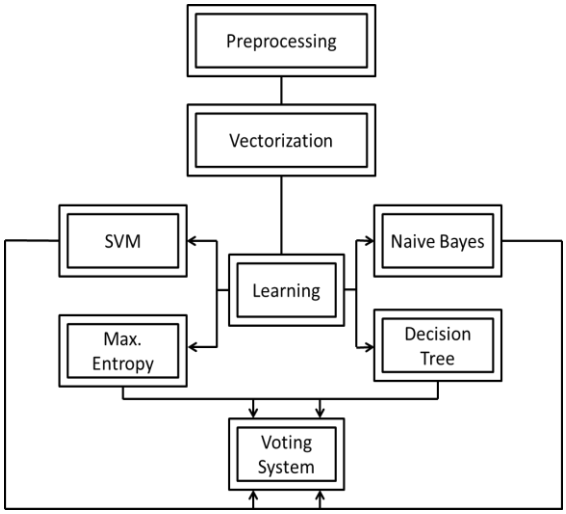

Figure 1: Proposed Approach

For each supervised classification methods used in this work, we identified three steps in the training phase: comment preprocessing, vectorization and learning.

Preprocessing: To make a correct comment preprocessing, we apply the following techniques:

- Elimination of symbols and special characters.
- Elimination of articles, adverbs, pronouns and prepositions (stopwords).
- Processing of hashtags.
- Correction of words with repeated letters.
- Filtration of words with ``@'' symbol as initial letter.
- Elimination of the characters ``RT''.
- URLs removal.
- Stemming of comments (opinions).

Vectorization: Each comment in the training data must be represented mathematically. There are different mathematical models to represent information. The most popular models are: boolean model, term frequency (TF), term frequency-inverse document frequency (TF-IDF) and Latent Semantic Analysis (LSA) (Codina, 2005). In this work, we decided to use the TF-IDF model to represent the comments of the corpus because it is more accurate and it has better results than the other models, (Salton and McGill, 1986). In Figure 2 it is shown an example of the corpus of tweets and its TF-IDF representation.

CORPUS OF TWEETS
Portada 'Público', viernes. Fabra al banquillo por 'orden' del Supremo; Wikileaks 'retrata' a 160 empresas espías.
Grande! RT @veronicacalderon "El periodista es alguien que quiere contar la realidad, pero no vive en ella"
Gonzalo Altozano tras la presentación de su libro 101 españoles y Dios. Divertido, emocionante y brillante
Mañana en Gaceta: TVE, la que pagamos tú y yo, culpa a una becaria de su falsa información sobre el cierre de @gaceta
Qué envidia "@mfcastineiras: Pedro mañana x la mañana me voy a Paris"
Más mañana en Gaceta. Amaiur depende de Uxue Barkos para crear grupo propio.

⇩

VECTOR SPACE REPRESENTATION OF TWEETS
(0,1),(1,1),(2,1),(3,1),(4,1),(5,1),(6,1),(7,1),(8,1),(9,1),(10,1),(11,2),(12,1),(13,1),(14,1)
(3,1),(9,1),(13,1),(15,1),(16,1),(17,1),(18,1),(19,1),(20,1)
(1,1),(2,1),(15,1),(21,1),(23,1),(24,1),(25,2),(26,2),(27,1),(28,1)
(16,1),(29,1),(30,1),(31,1),(32,1),(33,1)
(1,1),(11,1),(15,1),(29,1),(34,1),(35,1),(36,1),(37,1),(38,1)
(1,1),(2,1),(6,1),(11,2),(27,1),(29,1),(39,1),(40,1),(41,1),(42,1),(43,1),(44,1),(45,1)

⇩

TF-IDF WEIGHTING REPRESENTATION OF TWEETS
(0,031),(1,007),(2,0.12),(3,0.19),(4,0.31),(5,0.31),(6,0.19),(7,0.31),(8,0.31),(9,0.19, 10,0.31),(11,0.24),(12,0.31),(13,0.19),(14,0.31)
(3,0.26),(9,0.26),(13,0.26),(15,0.16),(16,0.26),(17,0.42),(18,0.42),(19,0.42),(20,0.42)
(1,0.07),(2,0.12),(15,0.12),(21,0.31),(22,0.31),(23,0.31),(24,0.31),(25,0.61),(26,0.31),(27,0.19),(28,0.31)
(16,0.29),(29,0.18),(30,0.47),(31,0.47),(32,0.47),(33,0.47)
(1,009),(11,0.16),(15,0.16),(29,0.16),(34,0.42),(35,0.42),(36,0.43),(37,0.43),(38,0.43)
(1,0.08),(2,0.13),(6,0.21),(27,0.21),(29,0.13),(39,0.34),(40,0.34),(41,0.34),(42,0.34),(43,0.34),(44,0.34),(45,0.34)

Figure 2: Vector Model Representation of Tweets

Learning: In this step, the classification algorithm receives as parameters the representative vectors of comments with their class labels. The class labels are: positive (P), negative (N), neutral (NEU) and none (NONE).

3.1.2 Classification

A classifier is a function that gives a discrete output, often denoted as class, to a particular input (Mohri, Rostamizadeh and Talwalkar, 2012). In this phase, the classifier receives a set of comments (the test data) and it evaluates this input to predict the corresponding class.

3.2 Our Proposal

In the first evaluation of the machine learning methods, the obtained accuracy results were slightly lower. For this reason, we propose to use two techniques to improve the results of classifiers. The first technique, called class reduction, removes one class label (NEU or NONE) with the aim of improving the margin of error of classifiers and reducing the number of classes to evaluate. The second technique, called naive voting, receives as input

parameters the optimized classifiers of the first technique. A more specific description of these techniques will be explained below.

3.2.1 Class Reduction

The basic idea of this technique was proposed in (Koppel and Schler, 2006). This technique is explained below.

Training and evaluation for three classes: We decided to train the classifiers considering three classes: Positive-Negative-Neutral and Positive-Negative-None. The classifiers were trained and tested in this way. The new results of classifiers using that simplification were better. Due to the improvement, we decided to join the partial results of these classifiers. With this union, we could classify the comments considering the four classes defined initially.

Union of partial results: We proposed to merge the partial results into single result. We established a set of rules to address the union of partial results of this class reduction technique, this rules are shown in Table 1.

Rule	Class Labels		Final Results
1	P	P	P
2	P	N	NEU
3	P	NONE	NONE
4	N	P	NEU
5	N	N	N
6	N	NONE	NONE
7	NEU	P	NEU
8	NEU	N	NEU
9	NEU	NONE	NONE

Table 1: Rules for Class Reduction Technique

3.2.2 Voting System

Our final technique presented in this paper was Voting System. We choose this method because all classifiers have a margin of error. Due to this margin of error, classifiers could classify incorrectly a comment. A voting system tries to reduce this margin of error. Voting systems are based on different classification methods. Many studies have used voting system to classify text. Kittler, Hatef and Matas (1998) and Kuncheva (2004) describe some of these methods. Rahman, Alam and Fairhurst (2002) show that in many cases the majority vote techniques are most efficient when classifiers are combined. Platie et al., (2009) and Tsutsumi, Shimada and Endo (2008) ensure that he following methods are the best voting systems for classification: Naive Voting, Weighted Voting, Maximun Choice Voting and F-Score/recall/precision Voting.

We proposed the Naive Voting technique, which has as input parameters the four classifiers proposed in this paper. Naive Voting is one of the simplest voting algorithms. In this technique, the comment is classified according to the majority agreement, i.e., the class with most votes in each classifier will be the winning class. The rules we applied for Naive Voting are described in Table 2.

Rule	Class Labels				Voting
	P	N	NEU	NONE	
1	4	0	0	0	P
2	3	0/1			P
3	2	0/1			P
4	0	4	0	0	N
5	0/1	3	0/1		N
6	0/1	2	0/1		N
7	0	0	4	0	NEU
8	0/1	3	0/1		NEU
9	0/1	2	0/1		NEU
10	0	0	0	4	NONE
11	0/1			3	NONE
12	0/1			2	NONE
13	2	2	0	0	P/N
14	2	0	2	0	NEU
15	2	0	0	2	NONE
16	0	2	2	0	NEU
17	0	2	0	2	NONE
18	0	0	2	2	NONE
19	2	0	2	0	NEU
20	1	1	1	1	NEU

Table 2: Naive Voting Rules

Each row of the Table 2 shows the votes obtained by each of the polarities (P-N-NEU-NONE) according to the output of the proposed classifiers. Due to we have 4 classifiers, the largest vote is 4 and the minimum is 0. Then, the class with the highest vote will be the winning class. In the event of a tie, a set of rules were established to determine the winning class. For example, in the case of a draw at 2 between positive and negative classes, a lottery was established to determine the winning. In other cases of a tie, it was chosen the NEU class or NONE class as the winner.

4 Experimental Results

4.1 Training and Test Data

We used the corpora provided by the organization of TASS 2015. For our purposes, we used the General Corpus and the Balanced General Corpus. The first one is composed of training and test set which contains 7219 and 60798 tweets, respectively. The Balanced General Corpus is a test subset which contains 1000 tweets only for test. A complete description of these corpora is explained in (Villena Román et al., 2015).

4.2 Evaluation of Classifiers

We performed a series of tests to address the Task 1 of TASS 2015, focusing on finding the global polarity of the Tweets corpora for 4 class labels (P-N-NEU-NONE). A general description of the "RUNs" that we have made for TASS 2015 are described in Table 3.

Tech.	Run-Id 60798	Run-Id 1000	Description
SVM	UCSP-RUN-2	UCSP-RUN-2	Support Vector Machine
NB	TestNB60000	UCSP-RUN-2-NB	Naive Bayes
ME	UCP-RUN-2-ME	TestME1000	Max. Entropy
DT	TestDT60000	TestDT10000	Decision Tree
SVM II	UCSP-RUN-1	UCSP-RUN-1	SVM + Class Reduction
NB II	UCSP-RUN-1-NB	UCSP-RUN-1-NB	NB + Class Reduction
ME II	UCSP-RUN-1-ME	UCSP-RUN-1-ME	ME + Class Reduction
DT II	UCSP-RUN-1-DT	UCSP-RUN-1-DR	DT + Class Reduction
Voting	UCSP-RUN-3	UCSP-RUN-3	Naive Voting

Table 3: Proposed Techniques

The results we have gotten for the evaluation of our proposal are shown in Table 4 (Evaluation of full test corpus) and Table 5 (Evaluation of 1k test corpus). It can be seen that class reduction techniques and our voting algorithm improve the accuracy of the original supervised classification algorithms.

Class reduction techniques improve results because they allow the classifier having to decide between fewer options and then the classifier could reduce its margin of error.

The voting algorithm gives good results because it takes into account the decisions of all the classifiers. This algorithm tries to reach a single decision that might be the best. A voting algorithm is like a consensus between all classifiers. But it is important to take into account that any voting algorithm is good as long as the majority of voters (classifiers) are good, otherwise, the voting algorithm will not have the expected results.

Approaches	Methods	Accuracy
Comparative	SVM	0.594
	NB	0.560
	ME	0.479
	DT	0.494
Proposal	SVM II	0.602
	NB II	0.560
	ME II	0.600
	DT II	0.536
	Voting	0.613

Table 4: Results of evaluating the Full Test Corpus

Approaches	Methods	Accuracy
Comparative	SVM	0.586
	NB	0.559
	ME	0.618
	DT	0.459
Proposal	SVM II	0.582
	NB II	0.636
	ME II	0.626
	DT II	0.495
	Voting	0.626

Table 5: Results of evaluating the 1k-Test Corpus

5 Conclusion

One of the main goals of this paper was to evaluate some supervised classification algorithms in the task of sentiment analysis. The results of evaluating the classifiers in initials experiments were not satisfactory. Using an optimization stage (class reduction and voting systems), accuracy improved

slightly compared to the original techniques. It could be shown that adequate voting algorithms improve the accuracy of classifiers. For proper operation of a voting system it is required to have multiple classifiers with a relatively high rate of efficiency. If a classifier fails, the other could give the correct prediction. But if most of classifiers give low results, then the voting system does not ensure a correct performance.

Acknowledgments

The research leading to the results has been founded by Programa Nacional de Innovación para la Competitividad y Productividad (Innóvate Perú)

References

Codina , L., 2005. Teoría de la recuperación de información: modelos fundamentales y aplicaciones a la gestión documental. *Revista internacional científica y profesional.*

Cui, H., V. Mittal and M. Datar. 2006. Comparative experiments on sentiment classification for online product reviews. *Proceedings of the 21st national conference on Artificial intelligence, Boston, Massachusetts.*

Hernandez, R. and Xiaoou Li. 2014. Sentiment analysis of texts in spanish based on semantic approaches with linguistic rules. *Proceedings of the TASS workshop at SEPLN 2014.*

Jimenez, S., E. Martinez, M. Valdivia and L. Lopez. 2014. SINAI-ESMA: An unsupervised approach for Sentiment Analysis in Twitter. *Proceedings of the TASS workshop at SEPLN 2014.*

Kittler, J., M. Hatef and J.Matas. 1998. On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence.*

Koppel, M. and J. Schler. 2006. The Importance of Neutral Examples for Learning Sentiment. *Dept. of Computer Science, Ramat Gan, Israel.*

Kuncheva, A. L.I. 2004. Combining Pattern Classifiers: Methods and Algorithms. *Jhon Wiley and Sons.*

Lopez, R., J. Tejada and M. Thelwall. 2012. Spanish Sentistrength as a Tool for Opinion Mining Peruvian Facebook and Twitter. *Artificial Intelligence Driven Solutions to Business and Engineering Problems.*

Mohri, M., A. Rostamizadeh and A. Talwalkar. 2012. Foundations of Machine Learning. *The MIT Press.*

Montejo-Raez, A., M.A. Garcia-Cumbreras and M.C. Diaz-Galiano. 2014. SINAI Word2Vec participation in TASS 2014. *Proceedings of the TASS workshop at SEPLN 2014.*

Pang, B., L. Lee and S. Vaithyanathan. 2002. Thumbs up?: sentiment classification using machine learning techniques. *Proceedings of the ACL-02 conference on Empirical methods in natural language processing, 10.*

Perea-Ortega, J. and A. Balahur. 2014. Experiments on feature replacements for polarity classification of Spanish tweets. *Proceedings of the TASS workshop at SEPLN 2014.*

Platie, M., M. Rouche, G. Dray and P. Poncelet. 2009. Is a voting approach accurate for opinion mining? *Centre de Recherche LGI2P.*

Rahman, A., H. Alam and M. Fairhurst. 2002. Multiple Classifier Combination for Character Recognition: Revisiting the Majority Voting System and Its Variation.

Salton G. and McGill M. 1986. Introduction to modern information retrieval.

San Vicente, I. and X. Saralegi. 2014. Looking for Features for Supervised Tweet Polarity Classification. *Proceedings of the TASS workshop at SEPLN 2014.*

Tsutsumi, K., K. Shimada and T. Endo. 2008. Movie Review Classification Based on a Multiple Classifier. *Department of Artificial Intelligence.*

Villena-Román, J., J. García-Morera, M. A. García-Cumbreras, E. Martínez-Cámara, M. T. Martín-Valdivia, and L. A. Ureña-López. 2015. Overview of TASS 2015.

Ye, Q., Z. Zhang and R. Law. 2009. Sentiment classification of online reviews to travel destinations by supervised machine learning approaches. *Expert Systems with Applications, 36:6527-6535*

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

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

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

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

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

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

1 Introduction

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

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

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

* We would like to express our gratitude for the financial support given by the Department of Software and Computer Systems at the University of Alicante, the Spanish Ministry of Economy and Competitivity (Spanish Government) by the project grants *ATTOS* (TIN2012-38536-C03-03) and *LEGOLANG* (TIN2012-31224), the European Commission by the project grant *SAM* (FP7-611312), and the University of Alicante by the project "*Explotación y tratamiento de la información disponible en Internet para la anotación y generación de textos adaptados al usuario*" (GRE13-15)

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

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

2 Related Work

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

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

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

3 Methodology

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

[1] www.daedalus.es/TASS2015

[2] www.socialisingaroundmedia.com

[3] wordnet.princeton.edu

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

3.1 Term Extraction

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

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

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

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

3.2 Term Scoring

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

[4]github.com/snowballstem

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

Figure 1: Term extraction process example

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

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

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

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

3.3 Learning

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

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

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

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

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

4 Evaluation

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

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

Table 1: TASS 2015 Official results

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

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

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

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

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

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

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

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

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

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

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

5 Conclusions

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

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

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

Bibliografía

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

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

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

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

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

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

Table 2: Datasets distribution

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

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

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

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

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

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

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

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

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

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

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

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

TASS 2015, septiembre 2015, pp 99-104 recibido 20-07-15 revisado 24-07-15 aceptado 28-07-15

Sentiment Classification using Sociolinguistic Clusters

Clasificación de Sentimiento basada en Grupos Sociolingüísticos

Souneil Park
Telefonica Research
souneil.park@telefonica.com

Resumen: Estudios sociolingüísticos sugieren una alta similitud entre el lenguaje utilizado por personas de una misma clase social. Análisis recientes realizados a gran escala sobre textos en Internet y mediante el uso de mineria, sustentan esta hipótesis. Datos como la clase social del autor, su geolocalización o afinidades políticas tienen efecto sobre el uso del lenguaje en dichos textos. En nuestro trabajo utilizamos la información sociolingüística del autor para la identificación de patrones de expresión de sentimiento. Nuestro enfoque expande el ámbito del analisis de textos al análisis de los autores mediante el uso de su clase social y afinidad política. Más concretamente, agrupamos tweets de autores de clases sociales o afinidades políticas similares y entrenamos clasificadores de forma independiente con el propósito de aprender el estilo lingüístico de cada grupo. Este mismo enfoque podría mejorarse en combinación con otras técnicas de procesado del lenguaje y aprendizaje automático.
Palabras clave: sociolingüística, clase social, estilo lingüístico, clustering de usuario.

Abstract: Sociolinguistic studies suggest the similarity of language use among people with similar social state, and recent large-scale computational analyses of online text are providing various supports, for example, the effect of social class, geography, and political preference on the language use. We approach the tasks of TASS 2015 with sociolinguistic insights in order to capture the patterns in the expression of sentiment. Our approach expands the scope of analysis from the text itself to the authors: their social state and political preference. The tweets of authors with similar social state or political preference are grouped as a cluster, and classifiers are built separately for each cluster to learn the linguistic style of that particular cluster. The approach can be further improved by combining it with other language processing and machine learning techniques.
Keywords: Sociolinguistics, Social Group, Linguistic Styles, User Clustering.

1 Introduction

The social aspect of language is an important means for understanding commonalities and differences in the language use as communication is inherently a social activity. Shared ideas and preferences of people are reflected in the language use, and frequently observed from various linguistic features such as memes, style, and word choices. The social aspect is also clear in the expression of sentiment, especially in social media. The social media platforms have many elements that encourage the use of similar expressions among social groups. For example, retweets and hashtags facilitate the adoption of expressions, and the short length of messages encourages the use of familiar expressions.

Our approach to the tasks of TASS 2015 (Villena-Román et al., 2015) is based on the insights of sociolinguistics. Specifically, we focus on the effect of social variables on linguistic variations; people who share similar preference or status may show similarity in the expression of sentiment than others. For each task, we cluster the tweets by people who share some social features (e.g., political orientation, occupation, or football team preference). In order to capture the style of the sociolinguistic clusters, a classification model is trained separately for each cluster.

While the primary benefit of the approach is that it can distinguish the different style of

sentiment expression among different social groups, it also mitigates the scale limitation of the training data. For instance, some football players of the Social TV corpus and some entity-aspect pair of the STOMPOL corpus have limited number of associated tweets. Clustering them with other tweets that are spoken by people with similar preference expands the amount of data that can be used for training.

The approach can be easily combined with other language processing and machine learning techniques. Since our approach mainly considers the characteristics of the authors rather than the text of tweets itself, combining it with more advanced language processing techniques complements each other. In addition, there is much room for future improvement as the current implementation of our approach uses primitive language processing methods due to the limited local Spanish knowledge of the author.

2 Related Work

The increasing availability of large-scale text corpora and the advances of big data processing platforms allows computational analysis of sociolinguistic phenomena. Many works in NLP and computational social science nowadays are taking the hypotheses of socio-linguistics as well as other social sciences and testing them with online data sets.

In the context of computational analysis of sociolinguistics theories, a number of works showed the effect of social features on linguistic variations. For example, Eisenstein et al. (2011) observed the difference in term frequency depending on the demographics and geographical information of people, and also that the different language use can play a significant role in predicting the demographics of authors. A similar study was conducted with the information about occupation (Preotiuc-Pietro et al., 2015), and gender (Wang et al., 2013). There are also works that specifically observed the relation between the expression of sentiment and social variables, for example, daily routine (Dodds et al., 2011) and urban characteristics (Mitchell et al., 2013).

The difference of the language use depending on the political/ideological preference has been explored as well. In the communication literature, researchers have conceptualized the phenomena as *framing* (Scheufele, 1999) and many studies analyzed how political and social issues are framed differently between media outlets and partisan organizations, and how they are related with the perception of the public. Many works are applying computational methods for similar purposes and observing the difference of language use from various online text data, for example, news articles (Park et al., 2011a), comments (Park et al., 2011b), and discussion forums (Somasundaran et al., 2009).

3 System Design

The classification systems that we have developed for the tasks share the central idea of using sociolinguistic clusters. We describe below the system developed for each task in order.

The classification tool is kept identical for all the tasks. We use linear SVM equipped with the Elastic Net regularizer as the classifier. Given a set of tweets, the system trains a binary classifier for each class in a one-vs-all manner and combines them for multi-class classification. The input text of the classifier goes through the TFIDF bag of words transformation. We optionally applied lemmatization and stop-word removal with FreeLing (Carreras et al., 2004) to the system for Task 1.

3.1 Task 1: General Sentiment Classification

The corpus of this task includes the tweets of selected famous people and information about them. The information about the people includes the occupation and political orientation.

Our system for this task clusters people based on their information, and uses the tweets of the clusters for training. The idea behind the system is that people with the same occupation or political orientation will have similar patterns in the expression of sentiments. A similar idea was tested with English tweets in Preotiuc-Pietro's work (2015), where they predicted the occupation of authors based on

their tweets. For example, journalists may have a certain way of expressing the sentiment, which can be different from that of celebrities.

We tested various clustering of people: clustering by the occupation, political orientation, and by both occupation and political orientation. The system trains a classifier for each cluster, only using the tweets made by the people of that cluster. Depending on the task granularity (5-level or 3-level), the system trains the classifiers accordingly.

3.2 Task 2 (a): Aspect-based Sentiment Analysis with SocialTV corpus

Unlike Task 1, the corpus does not have the information about the authors; thus, it is not clear how to cluster the tweets. However, the unique characteristic of the topic (the football match between Real Madrid and F.C. Barcelona) and the aspect-sentiment pair of the tweets provide useful implications about the authors. The rivalry between the two teams suggests that many of the authors prefer one of the two, and the aspect-sentiment pair gives hints about the preferred team. For example, if a tweet discusses Xavier Hernández and its sentiment is positive, it is possible to guess that the author prefers F.C. Barcelona, and the author will share the sentiment with other fans of F.C. Barcelona, who will commonly share the sentiment towards either F.C. Barcelona or Real Madrid.

Thus, we group the aspects based on the team affiliation. The players of each team are grouped as a single entity respectively, and one classifier is developed for each team. The rest of the aspects (e.g., Afición) are not clustered since they do not share a common membership with either of the teams. Classifiers are also developed separately for the rest of the aspects.

3.3 Task 2 (b): Aspect-based Sentiment Analysis with STOMPOL corpus

For this task, we cluster tweets in two levels. First, we cluster tweets by the entity-aspect pair. Thus, even if the tweets cover the same entity (party), they are treated to cover a different topic if the covered aspect is not the same. For example, a tweet about the economic proposal (aspect) of Podemos (entity) is distinguished from a tweet about the education policy (aspect) of Podemos (entity). It is also possible to cluster tweets only by entity; however, we consider both elements for clustering as all the tweets of the corpus have a specific aspect in association to the entity. In addition, it is also frequent that people evaluate a political party in multiple ways regarding different aspects; a person may evaluate the economic policies of Podemos positively but negatively its foreign policies. Theories of political communication, such as agenda setting and framing theory, suggest that people often recognize the parties and issues together when they evaluate the parties.

Second, we further cluster the tweets based on the characteristics of the political parties. For example, following the left vs. right dimension, the tweets about the entity Izquierda Unida and the aspect Economia are grouped with those about Podemos and Economia as the two parties would have similarity in terms of economic policies than other parties on the right wing. As a result, 10 clusters are produced (2 party groups *x* 5 aspects) and a classifier is developed separately for each cluster.

We compared two ways of grouping of the parties: first is the left vs. right dimension as in the example, and the second is the new vs. old dimension considering the new political landscape of Spain. The detail of the party grouping is shown in Table 1.

Left vs. Right		Old vs. New	
PSOE	PP	PP	Podemos
Podemos	Cs	PSOE	UPyD
IU		IU	Cs
UPyD			

Table 1: Two groupings of the parties

4 Results and Discussion

4.1 Task 1 General Sentiment Classification (5-levels, Full corpus)

For this task, we ran three versions of the method; first, clustering of the authors by occupation, second, by political orientation, third, by both. We submitted the first version (cluster by occupation) as it performed better

than the other two. The performance metrics are summarized in Table 2. The result and the performance trend were similar for the 1k test set corpus so we only describe the result of the full-corpus.

	Accuracy	Macro Average Precision	Macro Average Recall	Macro Average F-Measure
Occupation	0.462	0.385	0.450	0.415
Political Orientation	0.446	0.372	0.428	0.398
Both	0.423	0.351	0.401	0.374

Table 2. Performance Summary

The breakdown of the performance by sentiment category in Table 3 offers more insights. The performance for the category NEU and P is worse compared to that of other categories. While other optimization can be made for the two categories, we believe the method can be improved simply by having more number of examples of those categories in the training set. Compared to other categories, the current corpus includes much less examples for the two categories.

Category	P	R	F1
N	0.417	0.49	0.451
N+	0.35	0.539	0.424
NEU	0.064	0.217	0.099
NONE	0.659	0.405	0.502
P	0.094	0.554	0.161
P+	0.728	0.498	0.592

Table 3. Performance of Version 1 (Cluster-by-Occupation) by Sentiment Category

Interestingly, the performance further goes down when preprocessing (lemmatization and stopword removal) is conducted on the tweets. This performance drop was observed regardless of the version of our approach. The result suggests that conventional preprocessing removes important linguistic features that are relevant to sentiment expression. Due to the performance drop, we chose not to apply the preprocessing in the following tasks.

4.2 Task 1 General Sentiment Classification (3-levels, Full corpus)

We ran the same three versions of the method and the results are shown in Table 4. The performance is relatively higher than the 5-level classification task in general. Similar to the previous result, the version that clusters people by occupation performs better than the other two.

	Accuracy	Macro Average Precision	Macro Average Recall	Macro Average F-Measure
Occupation	0.594	0.518	0.491	0.504
Political Orientation	0.566	0.469	0.476	0.472
Both	0.549	0.454	0.459	0.457

Table 4. Performance Summary

The performance breakdown shows some difference from the previous task. First of all, the performance for the category P is much higher. We believe this is because the number of training examples of this category is higher than the previous task; the examples of P+ and P categories are merged together. We also see similar improvement for the category N. The category NEU still remains as a bottleneck. The improvement observed in the categories N and P suggests that similar improvement may be achieved for the category NEU if there are more examples in the training set.

Category	P	R	F1
N	0.53	0.835	0.648
NEU	0.11	0.098	0.104
NONE	0.808	0.226	0.353
P	0.625	0.806	0.704

Table 5. Performance of Version 1 (Cluster-by-Occupation) by Sentiment Category

4.3 Task 2a Aspect-based Sentiment Analysis with SocialTV corpus

As described, the approach to this task is to group the tweets by aspects that share the team membership in the training phase. The

performance of the approach is shown in Table 6.

	Accuracy	Macro Average Precision	Macro Average Recall	Macro Average F-Measure
Cluster-by-Team	0.631	0.460	0.484	0.471

Table 6. Performance Summary

Further analysis is required to understand the effect of the method. The breakdown of the performance by category does not show a clear pattern: while the tweets related to some players are identified very accurately but those of some other players are not; the performance does not differ much depending on the team of the players nor the sentiment expressed. We believe a larger test set that has enough samples for all players will better reveal the effect of the approach.

4.4 Task 2b Aspect-based Sentiment Analysis with STOMPOL corpus

Two versions of the approach are applied to the task: first, clustering the tweets of the same aspect by the parties of the same ideological leaning (left vs. right); second, by the novelty of the parties. The result is shown in Table 7.

	Accuracy	Macro Average Precision	Macro Average Recall	Macro Average F-Measure
Left vs. Right	0.557	0.252	0.297	0.272
New vs. Old	0.521	0.250	0.280	0.264

Table 7. Performance Summary

The version that groups by the ideological leaning of the parties performed better than the other version. The breakdown of the performance revealed that the approach performed better for the tweets that express a negative sentiment in general. For example, nine categories out of the top-10 categories in terms of F1 score were those expressing a negative sentiment. This is partly because many tweets related to politics often convey a negative sentiment hence there are more training examples with the negative sentiment.

5 Conclusion

In this paper, we present a sentiment classification method that utilizes sociolinguistic insights. The method is based on the idea that people with similar social state (e.g., occupation) or political orientation may show similarity also in the way they express their sentiment online. Thus, the method is focused on grouping authors with similar taste or occupation. A classifier is developed separately for each group to capture the similarities and differences of expression particularly within the group.

The method achieves around 0.45 and 0.6 in terms of accuracy for the 5-level Task 1 classification and 3-level Task 1 classification, respectively. It achieves 0.63 and 0.56 for the Social TV corpus and for the STOMPOL corpus. The result shows that the method performs better for the sentiment classes with more training examples. It can also be further improved by combining it with more language processing methods optimized to Spanish.

References

Carreras, Xavier, Isaac Chao, Lluis Padró, and Muntsa Padró. 2004. FreeLing: An Open-Source Suite of Language Analyzers. In *Proc. of LREC*.

Dodds, P. S., Harris, K. D., Kloumann, I. M., Bliss, C. A., & Danforth, C. M. 2011. Temporal patterns of happiness and information in a global social network: Hedonometrics and Twitter. *PloS ONE*, 6(12), e26752.

Eisenstein, J, Noah A. S., and Eric P. X. 2011. Discovering sociolinguistic associations with structured sparsity. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies.*

Mitchell, L., Frank, M. R., Harris, K. D., Dodds, P. S., & Danforth, C. M. 2013. The geography of happiness: Connecting twitter sentiment and expression, demographics,

and objective characteristics of place. *PLoS ONE* 8: e64417.

Park, S., Ko, M., Kim, J., Liu, Y., & Song, J. 2011a. The politics of comments: predicting political orientation of news stories with commenters' sentiment patterns. In *Proceedings of the ACM conference on Computer supported cooperative work.*

Park, S., Lee, K., & Song, J. 2011b. Contrasting opposing views of news articles on contentious issues. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies.*

Preotiuc-Pietro, D., Lampos, V., & Aletras, N. 2015. An analysis of the user occupational class through Twitter content. In *Proceedings of the 53th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies.*

Scheufele, D. A. 1999. Framing as a theory of media effects. *Journal of communication*, 49(1), 103-122.

Somasundaran, S., & Wiebe, J. 2009. Recognizing stances in online debates. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP. Association for Computational Linguistics.*

Villena-Román, J., García-Morera, J., García-Cumbreras, M.A., Martínez-Cámara, E., Martín-Valdivia, M. T., Ureña-López, L. A. 2015. Overview of TASS 2015. In *Proceedings of TASS 2015: Workshop on Sentiment Analysis at SEPLN*. CEUR-WS.org vol. 1397.

Wang, Y. C., Burke, M., Kraut, R. E. 2013. Gender, topic, and audience response: an analysis of user-generated content on facebook. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*.