A Survey of Multilingual Human-Tagged Short Message Datasets for Sentiment Analysis Tasks

Steiner-Correa, A.F.[1], Viedma-del-Jesus, M.I.[1], Lopez-Herrera, A.G.[2]

[1]Department of Market and Marketing Research, University of Granada (Spain),
{filipe.steiner@gmail.com, iviedma@ugr.es}

[2]Department of Computer Science and Artificial Intelligence, University of Granada (Spain),
{lopez-herrera@decsai.ugr.es}

Abstract

Today, the electronic Word of Mouth (eWOM) statements expressed on blogs, social media or shopping platforms are much frequent and enable customers to share his/her point of view about adquired products or services. These eWOM statements can be used for the industry to improve its products and services, and for customers for making better purchase decisions. Sentiment Analysis (SA) techniques can be used to extract and analyse these eWOM statements. Research in recent years on SA has advanced considerably and its applications in business management have grown exponentially. Automatic techniques (as machine learning, deep learning, statistic approaches and others) have been used for this purpose. However, training a machine for processing or analyzing sentiments is a hard task, mainly due to the complexity of the natural language. This task is more complicated in multilingual environments. There is still a great paucity regarding training datasets, one of the key resources in achieving more favourable results. Training datasets, in fact, are a reservoir of information serving to teach and refine the skills of automatic techniques. Hence, the higher the quality of the training datasets, the better predictive power of Sentiment Analysis tasks. English datasets are relatively easy to find in literature, however datasets in other languages are very scarce. So, this paper therefore describes and compiles information concerning 25 datasets gleaned from short messages (statements expressed in social media and shopping platforms) in seven different languages, for the most part from Twitter. For quality issues, all the resources were human-tagged, and they are currently available to the scientific community. A new sentiment dataset in English extracted from Twitter has also been drawn up and each message evaluated subjectively. The current survey therefore aims to provide essential quality information for future research related to automatic Sentiment Analysis in monolingual or multilingual scenarios.

Keywords: Sentiment Analysis; Dataset; Corpus; Short Messages, Multilingual, Twitter; Human-Tagged.

1. Introduction

The increasing use of Web for online activities like travel booking, e-commerce, social media communications, blogging, clicks streams, etc. enables to record, mine, filter, parse and summer large among information (referred in literature as big data). The big data is frequently composed by individual small portions of information. These small portions assume the form of short messages (texts, statements, opinions, etc.) expressed by customers with respect to acquired products or services, usually posted in social media (Twitter, Facebook, Instagram, etc.) or electronic shopping platforms (like Amazon, eBay, etc.), which are used by the industry to improve its products and services, and for customers for making better purchase decisions according to other consumers' experiences, this is referred in literature as electronic Word of Mouth (eWOM) (Hennig-Thurau et al., 2004). In recent years, the customers' favourite channel for expressing these eWOM statements is being Twitter.

The use of Twitter as a direct channel linking organisations and individuals that share recommendations, product reviews and services increases exponentially every day. This mode of communication provides all users the power to instantly and without obstacles create and share ideas. The more than 500 million registered Twitter users and more than 500 million daily Twitter messages[1] constitute a gold mine for companies as this application provides continuous essential data that can serve to improve their offer of products. Monitoring the network of Twitter messages therefore offers precious data as to what and how customers express themselves towards a company and towards its competition (Morinaga et al. 2002; Pang and Lee 2008).

1 http://www.internetlivestats.com/twitter-statistics/ [accessed July 14, 2017].

In this scenario, there is an increase in prominence of research and development of the techniques of Sentiment Analysis (SA) intended to sort short messages extracted from Twitter. Examples are the work of Gaspar et al. (2016), Go et al. (2009b), Lahuerta-Otero and Cordero-Gutiérrez (2016), Mukherjee and Bhattacharyya (2012), Spencer and Uchyigit (2012) and Yu and Wang (2015). To carry out studies to attain SA quality, it is fundamental to assemble corpuses or datasets. These are used in different tasks: Sentiment Classification (Serrano-Guerrero et al. 2015; Taboada 2016; Winkler et al. 2015; Ding et al. 2008; Sarvabhotla et al. 2011; Wilson et al. 2009), Subjectivity Analysis (Montoyo et al. (2012); Pang and Lee 2004; Wilson et al. 2009), Opinion Extraction (Sarvabhotla et al. 2011), the determination of viewpoints (Greene and Resnik 2009), identification of argumentation stands (Park et al. 2011), determination of emotional impact to events (Bernabé-Moreno 2015b and 2015c), geographically distributed event (Pino et al. 2016), the use in customers acquisition with marketing approaches (Bernabe-Moreno et al. 2015a; Wang et al. 2016), aspect-level based SA (Schouten et al. 2016), discurse style anaylisis (Nguyen and Jung 2017), etc. A more complete list de SA papers can be found in (Piryani et al. 2017).

SA usually is performed in a unique language or in a multilingual fashion as in (Boy and Moens 2009). SA also can be performed in different levels and/or granularity: documents, word, aspect, sentence, concept, phrase, clause, sense (Ravi and Ravi 2015). Analysis at the document-level identifies the sentiment as a whole and qualifies it as positive or negative. Phrase-level classification identifies the sentiment more specifically by determining the polarity (positive or negative) of each of a text's phrases. At the characteristic level it identifies the sentiments related with specific aspects of products, services or entities (Chen & Zimbra, 2010; Liu, 2012; Wilson, Wiebe, & Hoffmann, 2009). There are many applications linked to SA that can be divided into six large groups (Serrano-Guerrero, Olivas, Romero, & Herrera-Viedma, 2015):

A. Sentiment classification or sentiment polarity groups the information into three categories: positive, negative or neutral (Rushdi Saleh, Martín-Valdivia, Montejo-Ráez, & Ureña-López, 2011; Yu, Wu, Chang, & Chu, 2013). The information can be represented in numerical scales such as {-1,0,1} (the most common) indicative respectively of negative, neutral and positive sentiments, or [0 – 5] with zero tantamount to maximum negativity and 5 maximum to positivity (Li & Tsai, 2013; Martín-Valdivia, Martínez-Cámara, Perea-Ortega, & Ureña-López, 2013).

B. Subjectivity classification consists essentially in determining whether a sentence is subjective or objective. An objective sentence is relates facts and is usually easier to classify. A subjective sentence, in turn, tends to express other types of information as such as a personal belief, value or individual sentiment directly related to previous experiences. Certain authors see this task as a step prior to the classification of sentiments and affirm that a good classification at the subjective level potentiates the results of the classification of general sentiments (Barbosa & Feng, 2010; Esuli & Sebastiani, 2006; Montoyo, Martínez-Barco, & Balahur, 2012; Sarvabhotla, Pingali, & Varma, 2011).

C. Opinion summarization, according to Wang et al. (2013), consists essentially of identifying and extracting the main attributes and sentiments about an entity contained within one or several documents so as to detect opinions to identify relationships, characteristics and/or links (Beineke, Hastie, Manning, & Vaithyanathan, 2004; Pang & Lee, 2004).

D. Opinion retrieval, through two different types of evaluations linked to relevance and query, recover expressions/opinions in documents. This technique is commonly applied in document ranking (Lee, Song, Lee, Han, & Rim, 2012).

E. The sarcasm and irony approach consists in detecting sentences conveying these characteristics. This task, according to some authors, is the most difficult in SA since the definitions of sarcasm and irony are unclear (Reyes, Rosso, & Buscaldi, 2012; Farias Patti, Roso, 2016).

F. Other approaches following this line attempt to detect the gender of the author of the document (gender and authorship detection) (Montesi & Navarrete, 2008; Savoy, 2012; Seki, Kando, & Aono, 2009), or detect content with the objective of distorting public opinion about an entity (opinion spam detection) (Jindal & Liu, 2007; Ott, Choi, Cardie, & Hancock, 2011; Xie, Wang, Lin, & Yu, 2012).

Texts in general can also contain grammatical errors, abbreviations and colloquial expressions that complicate their classification (Balog, Mishne, & Rijke, 2006; Jindal & Liu, 2006). These require Natural Language Processing techniques (NLP) such as the following:

a) The negations technique that reverts sentiment so that its role be taken into account when analyzing the text (Abbasi, Chen, & Salem, 2008); b) POS or POS-tagging: (part of speech tagging) technique that identifies the adjectives and adverbs serving as indicators of sentiment (Turney, 2002); c) frequency and terms technique that considers the presence and frequency of unigrams or n-grams in the data (Dave, Lawrence, & Pennock, 2003; Pang, Lee, & Vaithyanathan, 2002), and d) opinion words and phrases, a technique widely applied to extract sentiments by either lexicon-based or statistical-based approaches (Hu & Liu, 2004). Other important resources must also be considered such as, among others, reducing the text to stem words and filtering of empty words (stop words) that help optimize data by streamlining so as to enhance classification.

Tools and online services as dictionaries have been proposed to facilitate some the analysis fhases as: *SentiWordNet* (Hung et al. 2013), *Senti-lexicon* (Kang et al. 2012), *SentiFul* (Neviarouskaya 2011) for word level SA; *EmotiNet* (Balahur et al. 2012) for emotion detection, *SenticNet* (Cambria et al. 2010), *SenticNet2* (Cambria et al. 2012) and *SenticNet3* (Cambria et al. 2014), (Poria et at. 2013) and (Tsai et at. 2013) for concept level SA; and *EmoSenticSpace* (Poria et al. 2014) for sense based SA. A more complete review of these tools and services can be found in (Ahmad et al. 2017).

Supervised and non-supervised approaches have been applied for automatic sentiment analysis. The more used in literature are Naïve Bayes, Super Vector Machine (SVM), decision tree, random forest and neural networks among others (Ravi and Ravi 2015).

The quality of the learning strongly depends on the quality of the corpuses. They can be largely divided in two groups: training and test. The training group comprises message assemblages evaluated subjectively and tagged manually indicating the sentiment or emotion content. This group serves to train the automated sentiment classifier (Parkhe and Biswas 2016; Roul et al. 2016; Jurafsky and Martin 2009; Nakov et al. 2013; Shamma et al. 2009). In this manner, the higher quality of the training dataset results in a higher precision in message classification. Once the classifier is trained, the test group is used to evaluate its efficiency. The test group usually consists of unsorted messages and regards the same topics treated in the training group.

The most common training datasets in the literature are classified by positive, negative and neutral polarity or represented by ranges using, for example, the number +5 to indicate very positive and -5 for very negative. Yet there are other less known corpuses that are tagged with other labels related to emotions including joy, anger, disgust and irrelevance (Go et al. 2009a; Román et al. 2015; Saif et al. 2012; Yu and Wang 2015).

Another important point regarding published training datasets is their scarcity in languages other than English (Saif et al. 2013). To meet this problem, this paper aims on the one hand to collect, identify and describe the data of 24 currently available, manually annotated, short messages (those posted in social media as Twitter or in electronic shopping platforms) corpuses of which ten are in English, four in Spanish (one in Mexican Spanish), four in the Portuguese (three in Brazilian Portuguese), one in Arabic and Jordanian Arabic, two in German, three in Italian and one in French. Furthermore, we advance an original dataset based on Twitter with a description of its characteristics, the method used to evaluate the messages and where it can be downloaded.

All these datasets can be used for future research in both monolingual and multilingual scenarios. All of these 25 datasets were human-tagged, and they are currently available to the scientific community. This paper revises them, summarizes their main characteristics, indicates in which context and with whose techniques are them used, and finally they are enhanced giving the readers links for downloading and disseminating.

This survey is therefore divided into four main parts: i) Section 2 describes the datasets according to language; ii) Section 3 introduces our dataset; iii) Section 4 summarizes the 25 datasets using a descriptive table; and iv) Section 5 draws the conclusions. We aspire to meet the main objectives behind this work focusing on gathering, describing and disseminating essential information so as to enrich and facilitate future research in the field of Sentiment Analysis.

2. Description of the datasets

This section describes 24 datasets, manually annotated for sentiment analysis tasks, of short messages corpuses in seven languages: English, Portuguese, Spanish, Arabic and Jordanian Arabic, German, Italian and French.

2.1. English Twitter Sentiment Datasets

This section describes the specifics each database according to English language, download address, authors and some of the scientific papers that have made reference to them.

2.1.1. *2000Entities*

The *2000Entities* corpus was published for the first time by Mukherjee et al. (2012) and used in (Mukherjee, Malu, Balamurali, and Bhattacharyya, 2012) to demonstrate the efficiency of the tool named TwiSent which uses the Naïve Bayes as the base algorithm and Spam Filter and Spell Checker algorithms to refine the data. It consists of 8,507 tweets regarding about 2,000 famous personalities from more than 20 different fields: movies, restaurants, television, politics, sports, education, philosophy, travel, books technology, banking and finance, business, music, environment, computers, automobiles, etc. The messages were tagged manually by four evaluators according to the following categories: positive, negative, objective-not-spam and objective-spam. The dataset can be consulted at http://www.mpi-inf.mpg.de/~smukherjee/data/twitter-data.tar.gz (accessed July 14, 2017).

2.1.2. Health Care Reform (HCR)

The *HCR* dataset of March 2010 comprised 2,516 tweets with reference to the healthcare reform hashtag "#hcr" (health care reform) introduced in 2010 by Barrack Obama (Speriosu et al. 2011). The messages were classified manually by five annotators in five different categories (positive, negative, neutral, irrelevant or other). In addition, the corpus was divided into tweets regarding training (839), development (838) and test (839) (Saif et al. 2013). This corpus has served in research conducted by Coletta et al. (2014), which used SVM classifier combined with a C3E-SL cluster ensemble, capable to combine classifier and cluster ensembles to refine the tweet analysis. Saif et al. (2012) presented a novel approach adding semantics as additional features into the training set and measure the correlation of the representative concept with negative/positive sentiment. Saif et al. (2014a,b) also exploit the semantic as additional feature. Speriosu et al. (2011) used a label propagation in a maximum entropy classifier trained on noisy labels and knowledge about word types encoded in a lexicon. Tsakalidis (2014) proposed an ensemble classifier that is trained on a general domain and adapts, on the desired (test) domain before classifying a document. The dataset can be consulted at https://bitbucket.org/speriosu/updown/downloads (accessed July 14, 2017).

2.1.3. Movies - UMICH SI650

The Movies UMICH SI650 dataset was designed by researchers at the University of Michigan between April and March 2011 for tasks related to SA. It has a training group comprising 7,086 tweets tagged manually as positive or negative. It also has a test group of 33,052 tweets. Both groups are associated with movies of different genres. Studies using this dataset include Dickinson et al. (2015) which applied a Word2Vec and Sent2Vec by a Deep Structured Sematic Model (DSSM) or the DSSM with convolutional-pooing structure (CDSSM) to form its vector representations and the Bag-of-Words model. Finally, Duncan and Zhang (2015) used a neural network for sentiment analysis tasks. Information relating to this resource and downloading can be found in at https://inclass.kaggle.com/c/si650winter11 (accessed July 14, 2017).

2.1.4. Obama-McCain Debate (OMD)

The *OMD* dataset was constructed from 3,238 tweets recorded during the first Obama and McCain presidential debate in September 2008. The messages were initially tagged by two annotators with a third playing the role of tiebreaker. The corpus consisted of 1,196 negative and 710 positive tweets. A total of 245 were considered mixed (Shamma et al. 2009; Mohammad et al. 2013). The dataset was used to evaluate different methods of supervised and unsupervised learning (Da Silva et al. 2014; Hu et al. 2013; Saif et al. 2012, 2013, 2014; Speriosu et al. 2011; Tsakalidis 2014). The dataset is available at https://bitbucket.org/speriosu/updown/downloads (accessed July 14, 2017).

2.1.5. Stanford

This corpus gathered by researchers at Stanford University comprises 16,000,000 tweets of which 800,000 are considered negative because they include the :(emoticon and 800,000 are considered positive because they contain the :) emoticon. A second classification was carried out manually generating a subset of 177 negative and 182 positive tweets. This corpus, published by Go et al. (2009a), has been widely used in the literature in its complete form, although a number of works chose to use its subjectively annotated reduced version for supervised and unsupervised machine learning works using a SVM, Nive Bayes, N-grams and others techniques (Bravo-Marquez et al. 2013; Hu et al. 2013; Saif et al. 2012, 2013, 2014; Speriosu et al. 2011; Tsakalidis 2014). This group of messages can be consulted in http://help.sentiment140.com/for-students (accessed July 14, 2017).

2.1.6. SemEval 2015 - Task11

This dataset was created for Task 11of the International SemEval (Semantic Evaluation) 2015 Sentiment Analysis of Figurative Language in Twitter workshop [2]. The event consisted of a round of evaluations or tasks designed to explore the meaning of language through computational semantic analysis systems. These tasks intended to provide new mechanisms to identify problems and solutions regarding complicated computations. The event also sought to articulate the dimensions related to the use of Natural Language Processing (NLP) (Rosenthal et al. 2015).

The SemEval 2015 - Task11 corpus therefore consists of a set of tweets using creative language rich in metaphor and irony. It is currently the only dataset that offers analysis of a high variety of tweets in figurative language. Specifically, it is divided into two groups: training and test. The training group has 8,000 messages assigned with polarity ranging from -5 to +5 (-5 being very negative and +5 very positive). The test group is smaller and consists of 1,000 tweets. This database is a fundamental resource in the research of Baca-Gomez et al. (2016) which was focused on a hybrid opinion mining approach and Ghosh et al. (2015) on the sentiment analysis of figurative language tasks using both a Cosine-similarity and a Mean-Squared-Error measure. It can be consulted at http://alt.qcri.org/semeval2015/task11/index.php?id=data-and-tools (accessed July 14, 2017).

2.1.7. Annotated-US2012-Election-Tweets

The US2012-Election dataset contains tweets from the months of August and September 2012 based on 21 hashtags alluding to the US presidential elections of 2011. Hashtag tweets containing the words Obama, Romney and Barack were also included in the dataset. After removing retweets and messages in other languages, a total of 170,000 tweets in English were retained. Classifying the messages from this dataset designed by Mohammad et al. (2015) was undertaken by crowdsource through Amazon's Mechanical Turk and CrowdFlower[3]. Two questionnaires called *HITS (human intelligence tasks)* were used to evaluate 2,042 tweets - each from different accounts - selected randomly and evaluated by native English speakers.

The first questionnaire served to determine the presence of emotions, the style and the purpose of the tweet. The goal was to determine the emotions of opposition by determining the tweets that show hypocrisy, mistakes, disagreement, ridicule, criticism and venting. A second goal was to identify the favourable tweets that show agreement, praise and support. The second questionnaire was hence a subgroup of the first. The 1,889 tweets among the first group considered emotional or containing emotional content were tagged according to eight basic emotions: trust, fear, surprise, sadness, disgust, anger, anticipation and joy.

---

2 http://alt.qcri.org/semeval2015/task11/ [accessed July 14, 2017].

3 https://crowdflower.com [accessed July 14, 2017].

These datasets are widely applied or mentioned in the following recent research: Cotelo et al. (2016) who do the Tweet categorization by combining content and structural knowledge; Fast et al. (2016) used this dataset on the Empath, a tool which draws connotations between words and phrases by deep learning a neural embedding and Mohammad et al. (2016) applied distant supervision techniques and word embeddings to further improve stance classification. Both the questionnaires and the datasets are intended for public use and can be consulted at http://saifmohammad.com (accessed July 14, 2017).

2.1.8. DAI-Labor English Dataset

This resource contains 7,200 tweets classified as either negative or positive. The messages were extracted solely from the network based on the emoticons :) :-) =) ;) :] :D ^-^ ^_^ indicating positive polarity and :(:-(:((-.- >:-(D: :/ indicating negative polarity. They therefore are not limited to any specific domain or topic. Each message was tagged by three annotators into three categories (positive, negative and neutral) by means of the Amazon Mechanical Turk tool. Concordance between the evaluations was established applying Fleiss' kappa coefficient (0.430) (Fleiss 1971). The dataset, assembled by Dai-Labor[4] with assistance from the Technical University of Berlin, was presented at the Workshop on Knowledge Discovery, Data Mining and Machine Learning (KDML-2012) (Narr et al. 2012) and used in supervised tasks mainly with Naïve Bayes. This resource and can be found in http://dainas.aot.tu-berlin.de/~andreas@dai/sentiment (accessed July 14, 2017).

2.1.9. Rateitall and Epinions Dataset

This dataset applied in Jakob & Gurevych (2010) and Wiegand & Klakow (2012), is annotated on the sentence and on the expression level distinguish between prior polarity and contextual polarity of a sentiment expression. Created by Toprak et al. (2010), it is a study of user-generated discourse on two levels of granularity: coarse-grained level and opinion on the topic, classified as explicit or polar facts. The corpus is composed of customer reviews (extracted from Rateitall.com and Epinions.com) on two different domains: online universities and online services. It embraces 240 university reviews (2786 sentences) and 234 service reviews (6091 sentences) considering the opinion expression at sentence-level from different aspects, such as polarity, strength, modifier, holder, and target. This resource is intended for public use and can be consulted at https://www.ukp.tu-darmstadt.de/data/sentiment-analysis/darmstadt-service-review-corpus/ (accessed July 14, 2017).

2.2. Portuguese Twitter Sentiment Datasets

This section describes the specifics each database according to Portuguese language, download address, authors and some of the scientific papers that have made reference to them.

2.2.1. *Notícias Globo* (Brazil)

This dataset contains 661 undifferentiated short messages. The data were extracted from the www.globo.com website and cover both Brazilian and international contexts. Messages were evaluated by two annotators with linguistic experience according to the categories of joy, disgust, fear, anger and sadness. This dataset, created and published in Dosciatti et al. (2013) and it was used in supervised tasks with SVM algorithm and stems from the *Emoções.BR* project, that studied sentiment analysis in Brazilian Portuguese texts. The dataset can be download at http://www.ppgia.pucpr.br/~paraiso/mineracaodeemocoes/ (accessed July 14, 2017).

2.2.2. *Política* (Brazil)

The *Política* (Politics) corpus, designed and applied in the work of Nascimento et al. (2015) that used language classifiers n-grams and Naïve Bayes. This resource contains 567 training tweets extracted between August and September 2011 regarding Brazil's political situation. The messages were classified by three different researchers either as positive, negative or neutral. Although the resource is not openly available, it can be accessed by contacting its creators.

2.2.3. *Entretenimento* (Brazil)

4 http://www.dai-labor.de/ [accessed July 14, 2017].

The *Entretenimento* (Entertainment) dataset, created and tested by the same group as *Política* (Nascimento et al. 2015) contains 384 training tweets extracted between August and September 2011 regarding Brazilian leisure and culture. The messages were tagged by three different investigators as negative (N), positive (P) and neutral (NEU). As in *Política* (cf. Section 2.2.2), the authors took special care to evaluate cases of irony and abbreviations using n-grams techniques. This resource, as in the case of the previous corpus, is also only available directly from its creators.

2.2.4. DAI-Labor Portuguese Dataset

This resource contains 18200 tweets classified as either negative or positive. The messages were extracted solely from the network based on the emoticons :) :-) =) ;) :] :D ^-^ ^_^ indicating positive polarity and :(:-(:((-.- >:-(D: :/ indicating negative polarity. They therefore are not limited to any specific domain or topic. Each message was tagged by three annotators into three categories (positive, negative and neutral) by means of the Amazon Mechanical Turk tool. Concordance between the evaluations was established applying Fleiss' kappa coefficient (0.408) (Fleiss 1971). The dataset, assembled by Dai-Labor [5] with assistance from the Technical University of Berlin, was presented at the Workshop on Knowledge Discovery, Data Mining and Machine Learning (KDML-2012) (Narr et al. 2012) and used it in supervised tasks mainly with Naïve Bayes. This resource can be found in http://dainas.aot.tu-berlin.de/~andreas@dai/sentiment (accessed July 14, 2017).

2.3. Spanish Twitter Sentiment Datasets

The Spanish language is represented in this study by five datasets. General-TASS, Social-TV-TASS and STOMPOL-TASS were designed and presented in the TASS 2014 (Román et al. 2015), a workshop to evaluate opinions in the Spanish language. Moreover, this event was a satellite of the annual conference of the Spanish Society for Natural Language Processing (SEPLN). These datasets can be accessed at http://www.sngularmeaning.team/TASS2015/tass2015.php#contact (accessed July 14, 2017). The main characteristic of the last two, SpanishCorpus3100 and SpanishCorpus1500, is that they pertain to Mexican Spanish.

2.3.1. General-TASS

The General-TASS dataset contains more than 68,000 tweets: 10% for training and 90% for test. The messages were extracted from Twitter between November 2011 and March 2012. The tweets refer to personalities and celebrities in the world of politics, economy, communications and culture. Although the context of extraction has a bias centred on Spain, the diverse nationalities of the authors (Spain, Mexico, Colombia, Puerto Rico, USA, among others) assures a global coverage throughout the Spanish-speaking world (Román et al. 2015). Training tweets were manually tagged into six categories: strong positive (P +), positive (P), neutral (NEU), negative (N), strong negative (N +) and one additional non sentiment tag (NONE). This dataset was used as a resource in many works as Perea-Ortega and Balahur (2014) and Vilares et al. (2014) in experiments based on feature replacements, deep learning techniques such as unsupervised pre-training, and sentiment-specific word embedding.

2.3.2. Social-TV-TASS

The Social-TV-TASS corpus was assembled during Spain's Copa del Rey Football Final on 16 April 2014 pitting Real Madrid against F.C. Barcelona. Tweets were extracted during the final 15 minutes of the match and the 15 minutes after the match. The dataset is composed of 1,773 training tweets and 1,000 test tweets classified manually into three categories: positive (P), neutral (NEU) and negative (N). The dataset was widely used by Hurtado and Pla (2014), Roncal et al. (2014) and Vilares et al. (2014) in works using n-grams analysis, deep learning and supervised SVM techniques.

2.3.3. STOMPOL-TASS

The STOMPOL (Spanish Tweets for Opinion Mining aspect at level about Politics) dataset consists of tweets extracted between 23 and 24 April 2014 regarding various political subjects in particular

5 http://www.dai-labor.de/ [accessed July 14, 2017].

economics and education. This dataset, widely referred to by Cumbreras et al. (2016), Park (2015) and Vilares et al. (2014), consists of 784 training tweets and 500 test tweets tagged manually by two different evaluators (and a third in the cases of disagreement) into three categories: positive (P), neutral (NEU) and negative (N). This data set was used in different tasks using a sociolinguistic clusters and deep learning techniques.

2.3.4. SpanishCorpus3100 - Mexican Spanish Dataset

Due to regional variations in the Spanish language, this paper also presents a dataset containing 3,100 postings in Mexican Spanish. The tagging, carried out by six individuals, divided the messages into two different series. Messages of the first series were evaluated in three categories: positive, neutral and negative. In the second round the same messages were divided into five categories: very positive, positive, neutral, negative and very negative. Consistency between the evaluations was assured by Krippendorff's alpha coefficient (Krippendorff 2004) resulting in two datasets serving as the main resources for their creators (Baca-Gomez et al. 2016). The dataset was used in hybrid opinion mining approaches which implements the Sequential Minimal Optimization (SMO) algorithm and can only be accessed from their creators.

2.4. Modern Standard Arabic (MSA) Twitter Dataset

This dataset includes messages in Modern Standard Arabic (MSA) and the regional Jordanian Arabic dialect. It was designed by Abdulla et al. (2013) and is widely applied in recent specialised literature (Al-Kabi et al. 2016; Al-Twairesh et al. 2015; Araujo et al. 2016; Obaidat et al. 2015). It was used in supervised and unsupervised tasks using a SVM, Naïve Bayes, Maximum Entropy, Bayes Net, and J48 algorithms. The resource consists of 2,000 tweets (1,000 positive and 1,000 negative) collected by Twitter Crawler[6]. Each tweet was classified by two experts with the intervention of a third in case of a tie. The corpus is strictly limited to two topics: politics and art. It can be obtained at https://archive.ics.uci.edu/ml/datasets/Twitter+Data+set+for+Arabic+Sentiment+Analysis# (accessed July 14, 2017).

2.5. German Twitter Sentiment Datasets

This section describes the specifics each database according to German language, download address, authors and some of the scientific papers that have made reference to them.

2.5.1. German Sentiment Dataset (GSD)

The German Sentiment Dataset (GSD) designed by Momtazi (2012) comprises 500 messages in German and is considered the first corpus of short messages in this language. The extracts come from different social media (i.e. Facebook and blogs) dealing with German celebrities from the world of music. Because of its pioneering character, the dataset has served as a reference in recent research (Scholz et al. 2012; Shalunts et al. 2014) used in several tasks of opinion mining as sentiment analysis, opinion extraction and the determination of viewpoints by methods as SVM, Naïve Bayes, Neural Net, Decision Tree and k-means. The dataset was annotated by three native German speakers using scales between 0 to -3 for negative evaluations and 0 to +3 for positive evaluations. Each message received two different classifications: polarity and strength. The concordance between the evaluations was carried out using two coefficients: Krippendorff's alpha (Krippendorff 2011) and Fleiss' kappa (Fleiss 1971). The dataset can be accessed at www.hpi.uni-potsdam.de/fileadmin/hpi/FG_Naumann/bachelorprojekte/BP2011N2/GermanSentimentData.zip (accessed July 14, 2017).

2.5.2. DAI-Labor German Dataset

This resource contains 1,800 tweets classified as either negative or positive. The messages were extracted solely from the network based on the emoticons :) :-) =) ;) :] :D ^-^ ^_^ indicating positive polarity and :(:-

6 http://www3.nd.edu/~dwang5/courses/spring15/assignments/A1/Assignment1_SocialSensing.html [accessed July 14, 2017].

(:((-.- >:-(D: :/ indicating negative polarity. They therefore are not limited to any specific domain or topic. Each message was tagged by three annotators into three categories (positive, negative and neutral) by means of the Amazon Mechanical Turk tool. Concordance between the evaluations was established applying Fleiss' kappa coefficient (0.419) (Fleiss 1971). The dataset, assembled by Dai-Labor[7] with assistance from the Technical University of Berlin, was presented at the Workshop on Knowledge Discovery, Data Mining and Machine Learning (KDML-2012) (Narr et al. 2012) and it was used in supervised tasks mainly with Naïve Bayes. This resource can be found in http://dainas.aot.tu-berlin.de/~andreas@dai/sentiment (accessed July 14, 2017).

2.6. Italian Twitter Sentiment Datasets

Of the three Italian datasets discussed in this paper, two (TWNews and TWSpino) are the result of the Senti-TUT1[8] project that sought to develop linguistic resources from the perspective of irony. Both corpuses described below focus on highly expressive and ironic political tweets. Designed by Bosco et al. (2015), they are widely referenced in the literature (V. Basile et al. 2014; Chafale and Pimpalkar 2014; Ravi and Ravi 2015). These datasets were applied to carry out SA experiments related to irony by sentiment classification at the message level on Italian tweets. It included three subtasks: subjectivity classification, polarity classification, and irony detection using the plutchik's wheel of emotions with fuzzy logic. Although these resources is no longer available on the Internet, it can be accessed directly from its authors.

The third corpus was designed for the SENTIPOLC task (SENTIment POLarity Classification) in the framework of EVALITA 2014 [9], a campaign aiming to evaluate natural language processing and voice tools in Italian. Its overall objective is to promote the development of Italian language technologies and voice tools where different systems and methods can be evaluated consistently. The SENTIPOLC task is divided into three subgroups: subjectivity classification, polarity classification and irony detection.

2.6.1. TWNews

The Italian tweets that make up this resource were collected from the Internet from 16 October 2011 to 3 February 2012 following the nomination of Mario Monti to replace Silvio Berlusconi as Prime Minister. A total of 3,228 unique messages were assembled by means of filters such as "mario monti/#monti", "governo monti/#monti", and "professor monti/#monti". The tagging process was carried out by five annotators (2 men and 3 women) who classified the tweets into five categories: POS (positive), NEG (negative), HUM (ironic), MIXED (positive and negative at the same time) and NONE (none of the above).

2.6.2 TWSpino

The TWSpino corpus is based on messages of the Twitter section of the Spinoza blog (http://www.spinoza.it (accessed July 14, 2017)) posted between July 2009 and February 2012. This popular Italian blog addresses political issues in a satirical tone. After deleting the tweets containing advertising (1.5%), a corpus of 1,159 unique messages was assembled. The classification process followed the same method as that of the TWNews dataset: five annotators (2 men and 3 women) who classified the tweets into five categories: POS (positive), NEG (negative), HUM (ironic), MIXED (positive and negative at the same time) and NONE (none of the above).

2.6.3. Sentipolc Task - Evalita 2014

The Sentipolc Task - Evalita 2014 corpus includes 7,410 randomly selected political (topic = 1) and generic (no specific domain) tweets (topic = 0) from the following items: "idtwitter" (twitter status id), "sbj" (subjectivity), "opos" (positive overall polarity), "oneg" (negative overall polarity), "iro" (irony), "lpos" (positive literal polarity), "lneg" (negative literal polarity), "top" (topic), "text" (twitter message). Each tweet was tagged by two expert annotators. A third annotator intervened in the cases of disagreement. This dataset, created by V. Basile et al. (2014) and widely applied in the literature (Basile

7 http://www.dai-labor.de/ [accessed July 14, 2017].

8 http://www.di.unito.it/ ~tutreeb/sentiTUT.html[accessed July 14, 2017].

9 http://www.evalita.it/2016/tasks/sentipolc[accessed July 14, 2017].

et al. 2015; Basile and Novielli 2014; Castellucci et al. 2014) combining lexicon and semantic features for a subjectivity classification, polarity classification and the pilot task irony detection through UNITOR system tool. Additional information can be downloaded at http://www.di.unito.it/~tutreeb/sentipolc-evalita16/data.html (accessed July 14, 2017).

2.7. DAI-Labor French Dataset

This resource contains 1,797 tweets classified as either negative or positive. The messages were extracted solely from the network based on the emoticons :) :-) =) ;) :] :D ^-^ ^_^ indicating positive polarity and :(:-(:((-.- >:-(D: :/ indicating negative polarity. They therefore are not limited to any specific domain or topic. Each message was tagged by three annotators into three categories (positive, negative and neutral) by means of the Amazon Mechanical Turk tool. Concordance between the evaluations was established applying Fleiss' kappa coefficient (0.244) (Fleiss 1971). The dataset, assembled by Dai-Labor[10] with assistance from the Technical University of Berlin, was presented at the Workshop on Knowledge Discovery, Data Mining and Machine Learning (KDML-2012) (Narr et al. 2012) and used in supervised tasks mainly with Naïve Bayes. This resource can be found in http://dainas.aot.tu-berlin.de/~andreas@dai/sentiment (accessed July 14, 2017).

3. Red Bull Twitter Sentiment Dataset (RSD)

The authors of this paper designed the Red Bull Twitter dataset from messages in English associated with the hashtag #givesyouwings, the hashtag that refers to the main publicity campaign of the Red Bull energy drink. The corpus contains two groups. The first is a training dataset consisting of 100 unique messages classified as positive, neutral or negative. The second, a test dataset, has 423 tweets extracted from the Internet during the months of May to June 2014.

To ensure the most precise subjective classification, we have adapted the Self-Assessment Manikin (SAM) scale (Hodes et al. 1985). This type of scale measures the degree of polarity through graphical representations of humanoid figures ranging from a happy face to a sad face. The figures were scored from 1 to 9 with 1 equal to a positive polarity and 9 equal to a negative polarity. Between the two extremes are intermediate scoring options (Fig1) (Bradley and Lang 1994).

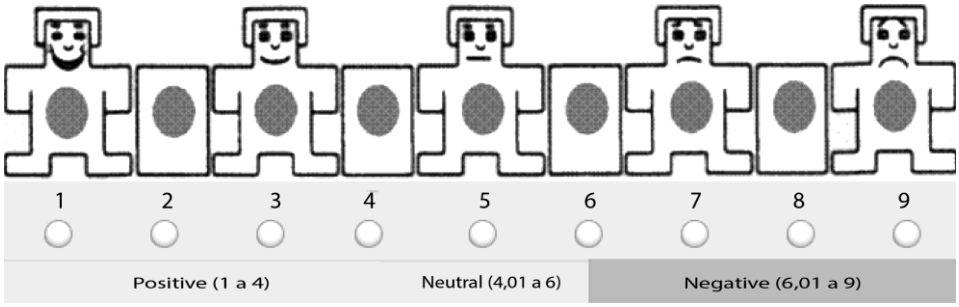


Fig. 1 Self-Assessment Manikin (SAM) evaluation scale developed by the authors from Bradley and Lang (1994).

The Red Bull dataset evaluations were conducted online by 152 students of the University of Granada (Spain) with knowledge of the brand. Specifically, the sampling consisted of women (64%) and men (36%). Most of those sampled individuals possessed a higher education and ranged in age between 18 and 44. Furthermore, 46% of the sample, in addition to having knowledge of the Red Bull brand, consumed the beverage. The dataset is available for free at http://mortero.ugr.es/steiner.

10 http://www.dai-labor.de/ [accessed July 14, 2017].

4. An overview of the different Twitter sentiment datasets

In this Section, we summarize the main features (language, domain, sort description and authors) of the 25 datasets revised in this p

Table I list the items of datasets reviewed in this paper. The table is divided into seven columns specifying the name of each datase topic covered by each resource, a brief description with the number of messages, the number of evaluators and filters, the classi (training and test) and finally references to the authors of each resource.

Table I Description of the different Twitter sentiment datasets.

Dataset	Language	Domain	Description	Class	Group	Courte
2000Entities	English	Movie, Restaurant, Television, Politics, Sports, etc.	8,507 tweets collected from a total of about 2000 different entities from 20 different domains	Positive, negative, objective-not-spam and objective-spam	Training	Mukher http://w [accesse
Health Care Reform (HCR)	English	Health care reform	2,516 tweets containing the hashtag "#hcr" (health care reform)	Positive, negative, irrelevant or other	Training / Test	Saif et a https://t [accesse
Movies - UMICH SI650	English	Movies	40,138 tweets compiled by the University of Michigan for a SA tasks	Positive and negative	Training / Test	Univers https://i [accesse
Obama-McCain Debate (OMD)	English	Politics	3,238 tweets extracted from U.S. presidential TV debate in September 2008	Positive, negative and mix or unknown	Training	Shamm https://t [accesse
Stanford Twitter Dataset	English	No domain	1,600,000 tweets based on emoticons	Positive and negative	Training / Test	Go et al http://h [accesse
SemEval_2015 Task 11	English	No domain	9,000 figurative tweets annotated with sentiment scores-ranging from -5...+5	Positive, neutral and negative	Training / Test	Rosenth http://al [accesse
Annotated-US2012-Election-Tweets	English	Politics	2,042 tweets annotated by 400 native English speakers extracted from August and September 2012	Trust, fear, surprise, sadness, disgust, anger, anticipation and joy.	Training	Moham http://sa [accesse
DAI-Labor English Dataset	English	No domain	7,200 tweets based on emoticons, annotated by 3 different researchers	Positive, neutral and negative	Training	Narr et http://da [accesse
RedBull Twitter Sentiment Dataset (RSD)	English	Beverage / Energy Drink	100 tweets annotated by 152 students, extracted in May 2014, based on the hashtag #givesyouwings from the RedBull company	Positive, neutral and negative	Training / Test	Steiner http://m [accesse
Rateitall.com /Epinions.com	English	Education and Services	Two different domains: 2786 sentences about 24 university	Polar fact, topic relevance, sentiment	Training	Toprak https://w

			reviews and 6091 sentences about service reviews collected from rateitall.com and epinions.com	target (anaphora resolution), sentiment expression (polarity/intensity), sentiment shifter and sentiment source		analysis [accesse
Noticias Globo	Portuguese (Brazil)	Feeds about Brazil and world	661 short messages extracted from www.globo.com	Joy, disgust, fear, anger and sadness	Training	Dosciat http://w [accesse
Política	Portuguese (Brazil)	Politics	567 tweets annotated by 3 different researchers	Positive, neutral and negative	Training	Nascim Availab
Entertainment	Portuguese (Brazil)	Entertainment, art and culture	384 tweets annotated by 3 different researchers	Positive, neutral and negative	Training	Nascim Availab
DAI-Labor Portuguese Dataset	Portuguese	No domain	1,800 tweets based on emoticons, annotated by 3 different researchers	Positive, neutral and negative	Training	Narr et http://da [accesse
General-TASS	Spanish	Politics, economics, communication and culture	68,000 tweets extracted from November 2011 until March 2012	6 classes: (P+) (P) (NEU) (N) (N+) and (NONE)	Training / Test	Román http://w [accesse
Social-TV-TASS	Spanish	Sport	2,773 tweets during the Spanish Copa del Rey Football Final between Real Madrid and F.C. Barcelona on 16 April 2014	Positive, neutral and negative	Training / Test	Román http://w [accesse
STOMPOL-TASS	Spanish	Politics	1,284 tweets extracted on the 23 and 24 April 2014	Positive, neutral and negative	Training / Test	Román http://w [accesse
SpanishCorpus3100	Mexican Spanish	No domain	3,100 tweets annotated by 6 different researchers	5 classes (P+) (P) (NEU) (N) (N+)	Training	Baca-G Availab
Modern Standard Arabic (MSA)	Arabic and Jordanian dialect	Politics and art	2,000 tweets annotated by 3 different researchers	2 Classes: Positive and Negative	Training	Assiri e https://a c+Senti [accesse
German Sentiment Dataset	German	German singers and musicians	500 short messages annotated by 3 native German speakers	Scores range -3...+3	Training	Momtaz www.hp potsdan N2/Ger [accesse
DAI-Labor German Dataset	German	No domain	1,800 tweets based on emoticons, annotated by 3 different researchers	Positive, neutral and negative	Training	Narr et http://da [accesse
TWNews	Italian	Politics	3,228 ironic tweets between the 16 October 2011 and 3 February 2012, annotated by 3 researchers	5 classes: POS, NEG, HUM, MIXED and NONE	Training	Bosco e Availab
TWSpino	Italian	Politics	1,159 ironic tweets extracted	5 classes: POS, NEG,	Training	Bosco e

			between July 2009 and February 2012, annotated by 3 researchers	HUM, MIXED and NONE		Availab
Sentipolc Task - Evalita 2014	Italian	Politics / Generic	7,410 ironic tweets annotated by 3 researchers	Subjectivity, positive overall polarity, negative overall polarity, irony, positive literal polarity, negative literal polarity	Training	V. Basi http://w [accesse
DAI-Labor French Dataset	French	No domain	1,797 tweets based on emoticons, annotated by 3 different researchers	Positive, neutral and negative	Training	Narr et http://da [accesse

5. Conclusions

The current social network of blogs, forums or wikis is a borderless channel of communication that serves as a platform for consumers to express their experiences/opinions of products and services (referred in literature as electronic Word of Mouth - eWOM) (Hennig-Thurau, Gwinner, Walsh, & Gremler, 2004). The extraction and classification of data from these social networks is therefore gaining ground day by day. Automatic Sentiment Analysis (SA) techniques emerge as a resource capable of collecting and classifying the data.

The quality of learning of the automatic SA techniques strongly depends on the quality of the corpuses used during the tuning process. The sentiment corpuses can be largely divided in two groups: training and test. The training group comprises message assemblages evaluated subjectively and tagged manually indicating the sentiment or emotion content. They indicate to the SA techniques which entries correspond to a certain class whereas the test groups are then used to check the quality of the classification. Hence, the higher the quality of the training datasets, the better predictive power of Sentiment Analysis tasks. English datasets are relatively easy to find in literature, however datasets in other languages are very scarce. Given the scarcity of these resources, this paper therefore describes and compiles information concerning 25 datasets gleaned from short messages (statements expressed in social media and shopping platforms) in seven different languages.

All these datasets can be used for future research in both monolingual and multilingual scenarios. All of these 25 datasets were human-tagged which guarantees the quality of them, and they are currently available to the scientific community. They were extracted mostly from the social network Twitter and recorded manually by scales of polarity or emotion. Some of the corpuses contain both training group and test group messages. The datasets are divided into seven languages: ten in English, four in Portuguese (three in the Brazilian Portuguese variant), four in Spanish (one language variant of Spanish of Mexico), one in Standard Arabic and Jordanian Arabic, two in German, three in Italian and one in French. In addition, they pertain to a variety of domains including politics, football, movies, health, product reviews and services.

We have also have created and put forward the Red Bull Twitter Sentiment Dataset (RSD). This resource comprises 100 training tweets evaluated subjectively by 152 students with a high education level, as well as 423 test tweets.

This paper revises all of them, summarizes their main characteristics, indicates in which context and with whose techniques are them used, and finally they are enhanced giving the readers links for downloading and disseminating.

To conclude, this paper presents a total of 1,778,081 training and test short messages. The total can be broken down into 1,681,618 in English, 75,157 in Spanish, 3,412 in Portuguese, 2,000 in Arabic, 2,300 in German, 11,797 in Italian, and 1,797 in French. These numbers enforce the main objective of this work which is to recompile vital information and resources for future research on automatic Sentiment Analysis.

References

Abbasi, A., Chen, H., & Salem, A. (2008). Sentiment analysis in multiple languages: Feature selection for opinion classification in Web forums. ACM Transactions on Information Systems, 26(3), 12:1-12:34. http://doi.org/10.1145/1361684.1361685.

Abdulla NA, Ahmed NA, Shehab MA, Al-Ayyoub M (2013) Arabic sentiment analysis: lexicon-based and corpus-based. In: Proceedings of IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT'13).

Ahmad, M., Aftab, S., Muhammad, S. S., & Waheed, U. (2017). Tools and Techniques for Lexicon Driven Sentiment Analysis: A Review. Int. J. Multidiscip. Sci. Eng, 8(1), 17-23.

Al-Kabi M, Al-Ayyoub M, Alsmadi I, Wahsheh H (2016) A prototype for a standard Arabic sentiment analysis corpus. International Arab Journal of Information Technology 13:163-170.

Al-Twairesh N, Al-Khalifa H, Al-Salman A (2015) Subjectivity and sentiment analysis of Arabic: trends and challenges. In: Proceedings of IEEE/ACS International Conference on Computer Systems and

Applications, (AICCSA'15), pp 148-155.
Araujo M, Pereira A, Reis J, Benevenuto F (2016) An evaluation of machine translation for multilingual sentence-level sentiment analysis. 1140-1145. doi: 10.1145/2851613.2851817.
Aryabhata K, Pingali P, Varma V (2011) Sentiment classification: a lexical similarity based approach for extracting subjectivity in documents. Information Retrieval 14:337-353. doi: 10.1007/s10791-010-9161-5.
Baca-Gomez YR, Martinez A, Rosso P, et al (2016) Web service SWePT: a hybrid opinion mining approach. Journal of Universal Computer Science 22:671-690.
Balahur A, Hermida JM and Montoyo A (2012) Building and exploiting EmotiNet, a knowledge base for emotion detection based on the appraisal theory model, IEEE Transaction on Affective Computing. 3:1, pp 88-101.
Balog, K., Mishne, G., & Rijke, M. De. (2006). Why are they excited? Identifying and explaining spikes in blog mood levels. In Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics: Posters & Demonstrations (EACL '06) (pp. 207–210). Retrieved from http://dl.acm.org/citation.cfm?id=1609010.
Barbosa, L., & Feng, J. (2010). Robust sentiment detection on twitter from biased and noisy data. Proceedings of the 23rd International Conference on Computational Linguistics: Posters. Association for Computational Linguistics, 36–44. Retrieved from http://dl.acm.org/citation.cfm?id=1944571.
Basile P, Basile V, Nissim M, Novielli N (2015) Deep tweets: from entity linking to sentiment analysis. In: Proceedings of Second Italian Conference on Computational Linguistics (CLiC-it'15), pp 41-45.
Basile P, Novielli N (2014) UNIBA at EVALITA 2014-SENTIPOLC Task: predicting tweet sentiment polarity combining micro-blogging, lexicon and semantic features. In: Proceedings of 4th Evaluation of NLP and Speech Tools for Italian (EVALITA'14), pp 58-63.
Basile V, Bolioli A, Nissim M, et al (2014) Overview of the Evalita 2014 sentiment polarity classification task. In: Proceedings of 4th Evaluation of NLP and Speech Tools for Italian (EVALITA'14), pp 50-57.
Beineke, P., Hastie, T., Manning, C., & Vaithyanathan, S. (2004). Exploring Sentiment Summarization. In Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text Theories and Applications (Vol. 7, pp. 1–4). Retrieved from http://www.aaai.org/Papers/Symposia/Spring/2004/SS-04-07/SS04-07-003.pdf.
Bernabé-Moreno, J., Tejeda-Lorente, A., Porcel, C., Fujita, H., & Herrera-Viedma, E. (2015a). CARESOME: A system to enrich marketing customers acquisition and retention campaigns using social media information. Knowledge-Based Systems, 80, 163-179.
Bernabé-Moreno, J., Tejeda-Lorente, A., Porcel, C., Fujita, H., & Herrera-Viedma, E. (2015b). Emotional Profiling of Locations Based on Social Media. Procedia Computer Science, 55, 960-969.
Bernabé-Moreno, J., Tejeda-Lorente, A., Porcel, C., & Herrera-Viedma, E. (2015c). A new model to quantify the impact of a topic in a location over time with Social Media. Expert Systems with Applications, 42(7), 3381-3395.
Boiy E, Moens MF (2009) A machine learning approach to sentiment analysis in multilingual Web texts. Informormation Retrieval 12, pp. 526–558, http://dx.doi.org/10.1007/s10791-008-9070-z.
Bosco C, Patti V, Bolioli A (2015) Developing corpora for sentiment analysis: the case of irony and Senti-TUT (extended abstract). In: Proceedings of Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI'15) 4158-4162. doi: 10.1109/MIS.2013.28.
Bradley MM, Lang PJ (1994) Measuring emotion: the self-assessment manikin and the semantic differential. Journal of Behavior Therapy and Experimental Psychiatry 25:49-59. doi: 10.1016/0005-7916(94)90063-9.
Bravo-Marquez F, Mendoza M, Poblete B (2013) Combining strengths, emotions and polarities for boosting twitter sentiment analysis. In: Proceedings of Second International Workshop on Issues of Sentiment Discovery and Opinion Mining (WISDOM'13), pp 1-9.
Cambria E, Speer R, Havasi C, Hussain A (2010) SenticNet: a publicly available semantic resource for opinion mining, in: AAAI Fall Symposium: Commonsense Knowledge, vol. 10, p. 02.
Cambria E, Havasi C, Hussain A (2012) SenticNet 2: a semantic and affective resource for opinion mining and sentiment analysis, Proc. 25th Int'l Florida Artificial Intelligence Research Society Conf., AAAI, pp. 202–207.
Cambria E, Olsher D, Rajagopal E (2014) SenticNet 3: a common and commonsense knowledge base for cognition-driven sentiment analysis, in: Twentyeighth AAAI Conference on Artificial Intelligence, pp. 1515–1521.
Castellucci G, Croce D, Cao D De, Basili R (2014) A multiple kernel approach for twitter sentiment analysis in Italian. In: Proceedings of 4th Evaluation of NLP and speech tools for Italian

(EVALITA'14), pp 98-103.
Chafale D, Pimpalkar A (2014) Review on developing corpora for sentiment analysis using plutchik's wheel of emotions with fuzzy logic. International Journal of Computer Sciences and Engineering (IJCSE) 2:14-18.
Chen, H., & Zimbra, D. (2010). AI and opinion mining. In IEEE Intelligent Systems (Vol. 25, pp. 74–76). http://doi.org/http://doi.org/10.1109/MIS.2010.75.
Coletta LFS, Silva NFF, Hruschka ER, Hruschka ERJ (2014) Combining classification and clustering for tweet sentiment analysis. In: Proceedings of Brazilian Conference on Intelligent Systems (BRACIS´14), pp 210-215.
Cotelo JM, Cruz FL, Enríquez F, Troyano JA (2016) Tweet categorization by combining content and structural knowledge. Information Fusion 31:54-64. doi: 10.1016/j.inffus.2016.01.002.
Cumbreras MÁG, Cámara EM, Román JV, Morera JG (2016) TASS 2015 - The evolution of the Spanish opinion mining systems. Procesamiento de Lenguaje Natural 56:33-40.
Da Silva NFF, Hruschka ER, Hruschka ERJ (2014) Tweet sentiment analysis with classifier ensembles. Decision Support Systems 66:170-179. doi: 10.1016/j.dss.2014.07.003.
Dave, K., Lawrence, S., & Pennock, D. M. (2003). Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In Proceedings of the 12th international conference on World Wide Web (pp. 519–528). http://doi.org/10.1145/775152.775226.
Dickinson B, Ganger M, Hu W (2015) Dimensionality reduction of distributed vector word representations and emoticon stemming for sentiment analysis. Journal of Data Analysis and Information Processing 3:153-162. doi: 10.4236/jdaip.2015.34015.
Ding X, Liu B, Yu PS (2008) A holistic lexicon-based approach to opinion mining. In: Proceedings of International conference on Web search and web data mining (WSDM'08), pp 231-239.
Dosciatti MM, Ferreira LPC, Paraiso EC (2013) Identificando emoções em textos em português do Brasil usando máquina de vetores de suporte em solução multiclasse. In: Proceedings of X Encontro nacional de inteligência artificial e computacional.
Duncan B, Zhang Y (2015) Neural networks for sentiment analysis on twitter. In: Proceedings of 14th International conference on cognitive informatics & cognitive computing (ICCI'CC'15), pp 275-278.
Esuli, A., & Sebastiani, F. (2006). Determining term subjectivity and term orientation for opinion mining. In Proceedings of the 11th Meeting of the European Chapter of the Association for Computational Linguistics (EACL-2006) (Vol. 2, pp. 193–200). Retrieved from http://acl.ldc.upenn.edu/eacl2006/main/papers/13_1_esulisebastiani_192.pdf.
Farías, D. I. H., Patti, V., & Rosso, P. (2016). Irony detection in Twitter: The role of affective content. ACM Transactions on Internet Technology (TOIT), 16(3), 19.
Fast E, Chen B, Bernstein MS (2016) Empath: understanding topic signals in large-scale text. In: Conference on human factors in computing systems (CHI'16), pp 4647-4657.
Fleiss JL (1971) Measuring nominal scale agreement among many raters. 76:378-382.
Gaspar R, Pedro C, Panagiotopoulos P, Seibt B (2016) Beyond positive or negative: qualitative sentiment analysis of social media reactions to unexpected stressful events. Computers in Human Behavior 56:179-191. doi: 10.1016/j.chb.2015.11.040.
Ghosh A, Li G, Veale T, et al (2015) SemEval-2015 Task 11: Sentiment analysis of figurative language in twitter. In: Proceedings of 9th International Workshop on Semantic Evaluation (SemEval'15), pp 470-478.
Go A, Bhayani R, Huang L (2009a) Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford, 2009 5. doi: 10.1016/j.sedgeo.2006.07.004.
Go A, Huang L, Bhayani R (2009b) Twitter sentiment analysis. CS224N - Final Project Report 17. doi: 10.1007/978-3-642-35176-1_32.
Greene S, Resnik P (2009) More than words: syntactic packaging and implicit sentiment. In: Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the ACL, pp 503-511.
Hennig-Thurau, T., Gwinner, K. P., Walsh, G., & Gremler, D. D. (2004). Electronic word-of-mouth via consumer-opinion platforms: What motivates consumers to articulate themselves on the Internet? Journal of Interactive Marketing, 18(1), 38–52. http://doi.org/10.1002/dir.10073.
Hodes RL, Cook EW, Lang PJ (1985) Individual differences in autonomic response: conditioned association or conditioned fear? Psychophysiology 22:545-560. doi: 10.1111/j.1469-8986.1985.tb01649.x.
Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining (KDD'04) (pp. 168–177). http://doi.org/http://dx.doi.org/10.1145/1014052.1014073

Hu X, Tang L, Tang J, Liu H (2013) Exploiting social relations for sentiment analysis in microblogging. In: Proceedings of Sixth ACM International Conference on Web Search and Data Mining (WSDM'13), pp 537-546.

Hung C, Lin HK, Yuan C (2013) Using objective words in SentiWordNet to improve word-of-mouth sentiment classification, IEEE Transactions on Intelligent Systems 2, pp 47–54.

Hurtado L-F, Pla F (2014) ELiRF-UPV en TASS 2014: análisis de sentimientos, detección de tópicos y análisis de sentimientos de aspectos en Twitter. Procesamiento del Lenguaje Natural pp 1-7.

Kang H, Yoo SJ, Han D (2012) Senti-lexicon and improved Naïve Bayes algorithms for sentiment analysis of restaurant reviews, Expert Systems with Applications 39 pp 6000–6010.

Jakob, N., Gurevych, I. (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 (pp. 1035–1045). Retrieved from http://portal.acm.org/citation.cfm?id=1870759.

Jindal, N., & Liu, B. (2006). Identifying comparative sentences in text documents. In Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR'06) (p. 244). http://doi.org/10.1145/1148170.1148215.

Jindal, N., & Liu, B. (2007). Review spam detection. In Proceedings of WWW-2007 (pp. 1189–1190). http://doi.org/10.1145/1242572.1242759.

Jurafsky D, Martin JH (2009) Speech and language processing: an introduction to natural language processing.

Krippendorff K (2004) Content analysis: an introduction to its methodology, 2nd edn.

Krippendorff K (2011) Computing Krippendorff's alpha-reliability. Departmental Papers (ASC) 1-12.

Lahuerta-Otero E, Cordero-Gutiérrez R (2016) Looking for the perfect tweet. the use of data mining techniques to find influencers on Twitter. Computers in Human Behavior 64:575–583. doi: 10.1016/j.chb.2016.07.035.

Lee, S. W., Song, Y. I., Lee, J. T., Han, K. S., & Rim, H. C. (2012). A new generative opinion retrieval model integrating multiple ranking factors. Journal of Intelligent Information Systems, 38(2), 487–505. http://doi.org/10.1007/s10844-011-0164-5.

Li, S.-T., & Tsai, F.-C. (2013). A fuzzy conceptualization model for text mining with application in opinion polarity classification. Knowledge-Based Systems, 39, 23–33. http://doi.org/10.1016/j.knosys.2012.10.005.

Liu, B. (2012). Sentiment Analysis and Opinion Mining. Synthesis lectures on human language technologies (Vol. 5). Morgan & Claypool Publishers. http://doi.org/10.2200/S00416ED1V01Y201204HLT016.

Martín-Valdivia, M. T., Martínez-Cámara, E., Perea-Ortega, J. M., & Ureña-López, L. A. (2013). Sentiment polarity detection in Spanish reviews combining supervised and unsupervised approaches. Expert Systems with Applications, 40(10), 3934–3942. http://doi.org/10.1016/j.eswa.2012.12.084.

Mohammad SM, Kiritchenko S, Zhu X (2013) NRC-Canada: building the state-of-the-art in sentiment analysis of tweets. In: Proceedings of seventh international workshop on semantic evaluation exercises (SemEval'13), pp 321-327. arXiv preprint arXiv:1308.6242. Accessed 09 Nov 2016.

Mohammad SM, Sobhani P, Kiritchenko S (2016) Stance and sentiment in tweets. ACM Transactions on Embedded Computing Systems 0:22. arXiv preprint arXiv:1605.01655v1. Accessed 09 Nov 2016.

Mohammad SM, Zhu X, Kiritchenko S, Martin J (2015) Sentiment, emotion, purpose, and style in electoral tweets. Information Processing and Management 51:480-499. doi: 10.1016/j.ipm.2014.09.003.

Momtazi S (2012) Fine-grained German sentiment analysis on social media. In: Proceedings of 9th Intl. Conf. on Language Resources and Evaluation, pp 1215-1220.

Montoyo A, Martínez-Barco P, Balahur A (2012) Subjectivity and sentiment analysis: an overview of the current state of the area and envisaged developments. Decision Support Systems 53:675–679. doi: 10.1016/j.dss.2012.05.022.

Montesi, M., & Navarrete, T. (2008). Classifying web genres in context: A case study documenting the web genres used by a software engineer. Information Processing and Management, 44(4), 1410–1430. http://doi.org/10.1016/j.ipm.2008.02.001.

Morinaga S, Yamanishi K, Tateishi K, Fukushima T (2002) Mining product reputations on the web. In: Proceedings of Eighth ACM SIGKDD international conference on Knowledge discovery and data mining (KDD'02), pp 341-349.

Mukherjee S, Bhattacharyya P (2012) Sentiment analysis in twitter with lightweight discourse analysis. In: Proceedings of Coling, pp 1847-1864.

Mukherjee S, Malu A, Balamurali AR, Bhattacharyya P (2012) TwiSent: a multistage system for

analyzing sentiment. In: Proceedings of Conference on Information and Knowledge Management (CIKM'12), pp 2531-2534.
Nakov P, Rosenthal S, Kozareva Z, et al (2013) SemEval-2013 Task 2: sentiment analysis in twitter. In: Proceedings of International Workshop on Semantic Evaluation (SemEval'13), pp 312-320.
Narr S, Hülfenhaus M, Albayrak S (2012) Language-independent twitter sentiment analysis. In: Proceedings of Knowledge Discovery and Machine Learning (KDML'12), pp 12-14.
Nascimento P, Aguas R, Lima D de, et al (2015) Análise de sentimento de tweets com foco em notícias. Revista Eletrônica de Sistemas de Informação 14:12. doi: 10.5329/RESI.
Neviarouskaya A, Prendinger H, Ishizuka M (2011) SentiFul: a lexicon for sentiment analysis, IEEE Transactions on Affective Computing 2:1 pp.
Nguyen, H. L., & Jung, J. E. (2017). Statistical approach for figurative sentiment analysis on social networking services: a case study on twitter. *Multimedia Tools and Applications*, *76*(6), 8901-8914.
Obaidat I, Mohawesh R, Al-Ayyoub M, et al (2015) Enhancing the determination of aspect categories and their polarities in Arabic reviews using lexicon-based approaches. In: Proceedings of Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT'15), pp 1-6.
Ott, M., Choi, Y., Cardie, C., & Hancock, J. T. (2011). Finding deceptive opinion spam by any stretch of the imagination. In Proceedings of the 49th Annua Meeting of the Association for Computational Linguistics (pp. 1–11). Retrieved from http://arxiv.org/abs/1107.4557.
Pang B, Lee L (2004) A sentimental education: sentiment analysis using subjectivity summation based on minimum cuts. In: Proceedings of 42nd Annual Meeting on Association for Computational Linguistics (ACL'04), pp 271-278. doi: 10.3115/1218955.1218990.
Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. In Conference on Empirical Methods in Natural Language Processing (EMNLP'02) (pp. 79–86). http://doi.org/10.3115/1118693.1118704.
Pang, B., & Lee, L. (2004). A sentimental education: Sentiment analysis using subjectivity summation based on minimum cuts. In Proceedings of 42nd Annual Meeting on Association for Computational Linguistics (ACL´04) (pp. 271–279). http://doi.org/10.3115/1218955.1218990.
Pang B, Lee L (2008) Opinion mining and sentiment analysis. Foundations and trends in information retrieval 2:1-135. doi: http://dx.doi.org/10.1561/1500000011.
Park S (2015) Sentiment classification using sociolinguistic clusters. In: Proceedings of TASS 2015: Workshop on Sentiment Analysis at SEPLN, pp 99-104.
Park S, Lee K, Song J (2011) 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 (HLT'11), pp 340-349.
Parkhe V, Biswas B (2016) Sentiment analysis of movie reviews: finding most important movie aspects using driving factors. Soft Computing 20:3373-3379. doi: 10.1007/s00500-015-1779-1.
Perea-Ortega JM, Balahur A (2014) Experiments on feature replacements for polarity classification of Spanish tweets. In: Proceedings of TASS 2014: Workshop on Sentiment Analysis at SEPLN, pp 1-7.
Pino, C. Kavasidis, I, Spampinato, C. (2016). GeoSentiment: a Tool for Analyzing Geographically Distributed Event-related Sentiments. 2016 In: Proceedings of 13th IEEE Annual Consumer Communications & Networking Conference (CCNC).
Piryani, R., Madhavi, D., & Singh, V. K. (2017). Analytical mapping of opinion mining and sentiment analysis research during 2000–2015. *Information Processing & Management*, *53*(1), 122-150.
Poria S, Gelbukh A, Hussain A, Howard N, Das D, Bandyopadhyay S (2013) Enhanced SenticNet with affective labels for concept-based opinion mining, IEEE Transactions on Intelligent Systems 2 pp 31–38.
Ravi K, Ravi V (2015) A survey on opinion mining and sentiment analysis: Tasks, approaches and applications. Knowledge-Based Systems 89:14-46. doi: 10.1016/j.knosys.2015.06.015.
Reyes, A., Rosso, P., & Buscaldi, D. (2012). From humor recognition to irony detection: The figurative language of social media. Data & Knowledge Engineering, 74, 1–12.
Román JV, Morera JG, Cámara EM, Zafra SMJ (2015) TASS 2014 - The challenge of aspect-based sentiment analysis. Procesamiento de Lenguaje Natural 54:61-68.
Roncal ISV, Urizar XS (2014) Looking for features for supervised tweet polarity classification. In: Proceedings of TASS 2014: Workshop on Sentiment Analysis at SEPLN.
Rosenthal S, Nakov P, Kiritchenko S, et al (2015) Semeval-2015 task 10: sentiment analysis in twitter. In: Proceedings of 9th International Workshop on Semantic Evaluation (SemEval'15), pp 451-463.
Roul RK, Asthana SR, Kumar G (2016) Study on suitability and importance of multilayer extreme learning machine for classification of text data. Soft Computing 1-18. doi: 10.1007/s00500-016-2189-8.

Rushdi Saleh, M., Martín-Valdivia, M. T., Montejo-Ráez, A., & Ureña-López, L. A. (2011). Experiments with SVM to classify opinions in different domains. Expert Systems with Applications, 38(12), 14799–14804. http://doi.org/10.1016/j.eswa.2011.05.070.

Sarvabhotla, K., Pingali, P., & Varma, V. (2011). Sentiment classification a lexical similarity based approach for extracting subjectivity in documents. *Information Retrieval*, *14*(3), 337–353.

Saif H, Fernandez M, He Y, Alani H (2013) Evaluation datasets for Twitter sentiment analysis: a survey and a new dataset, the STS-Gold. In: Proceedings of 1st International Workshop on Emotion and Sentiment in Social and Expressive Media: Approaches and Perspectives from AI (ESSEM'13), pp 9-21.

Saif H, He Y, Alani H (2012) Semantic sentiment analysis of twitter. In: Proceedings of The 11th International Semantic Web Conference (ISWC'12), pp 508-524.

Saif H, He Y, Fernandez M, Alani H (2014a) Adapting sentiment lexicons using contextual semantics for sentiment analysis of Twitter. In: Proceedings of European Semantic Web Conference (ESWC'14), pp 54-63.

Saif H, He Y, Fernandez M, Alani H (2014b) Semantic patterns for sentiment analysis of twitter. In: Proceedings of Proceedings of the 13th International Semantic Web Conference - Part II (ISWC'14), pp 324-340.

Savoy, J. (2012). Authorship attribution based on specific vocabulary. ACM Transactions on Information Systems, 30(2), 1–30. http://doi.org/10.1145/2180868.2180874.

Seki, Y., Kando, N., & Aono, M. (2009). Multilingual opinion holder identification using author and authority viewpoints. Information Processing and Management, 45(2), 189–199. http://doi.org/10.1016/j.ipm.2008.11.004.

Serrano-Guerrero, J., Olivas, J. A., Romero, F. P., & Herrera-Viedma, E. (2015). Sentiment analysis: a review and comparative analysis of web services. Information Sciences, 311, 18-38.

Schouten, K., & Frasincar, F. (2016). Survey on aspect-level sentiment analysis. IEEE Transactions on Knowledge and Data Engineering, 28(3), 813-830.

Scholz T, Conrad S, Hillekamps L (2012) Opinion mining on a German corpus of a media response analysis. In: Proceedings of International Conference on Text, Speech and Dialogue, pp 39-46.

Shalunts G, Backfried G, Prinz K (2014) Sentiment analysis of German social media data for natural disasters. In: Proceedings of 11th International conference on information systems for crisis response and management (ISCRAM'14), pp 752-756.

Shammas DA, Kennedy L, Churchill EF (2009) Tweet the debates: understanding community annotation of uncollected sources. In: Proceedings of The first SIGMM workshop on Social media (WSM'09), pp 1-8.

Spencer J, Uchyigit G (2012) Sentimentor: sentiment analysis of twitter data. In: Proceedings of The 1st International Workshop on Sentiment Discovery from Affective Data (SDAD'12), pp 56-66.

Speriosu M, Sudan N, Upadhyay S, Baldridge J (2011) Twitter polarity classification with label propagation over lexical links and the follower graph. In: Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP'11), pp 53-63.

Taboada, M. (2016). Sentiment Analysis: An Overview from Linguistics. Annual Review of Linguistics, 2, 325-347.

Toprak, C., Jakob, N., Gurevych, I. (2010). Sentence and Expression Level Annotation of Opinions in User-Generated Discourse. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (Vol. 1, pp. 575–584). Retrieved from http://www.aclweb.org/anthology/P10-1059.

Tsai ACR, Wu CE, Tsai RTH, Hsu JYJ (2013) Building a concept-level sentiment dictionary based on commonsense knowledge, IEEE Transactions on Intelligent Systems 2, pp 22–30.

Tsakalidis A, Papadopoulos S, Kompatsiaris I (2014) An ensemble model for cross-domain polarity classification on Twitter. In: Conference on Web Information Systems Engineering - Part II (WISE'14), pp 168-177.

Turney, P. D. (2002). Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL'02), (July), 417–424. http://doi.org/10.3115/1073083.1073153.

Vilares D, Alonso MA (2016) A review on political analysis and social media. Procesamiento de Lenguaje Natural 56:13-24.

Vilares D, Doval Y, Alonso MA, Gómez-Rodríguez C (2014) LyS at TASS 2014: a prototype for extracting and analysing aspects from Spanish tweets. In: Proceedings of TASS 2014: Workshop on Sentiment Analysis at SEPLN.

Vilares D, Doval Y, Alonso MA, Gómez-Rodríguez C (2015) LyS at TASS 2015: deep learning

experiments for sentiment analysis on Spanish tweets. In: Proceedings of TASS 2015: Workshop on Sentiment Analysis at SEPLN, pp 47-52.
Wang, D., Zhu, S., & Li, T. (2013). SumView: A Web-based engine for summarizing product reviews and customer opinions. Expert Systems with Applications, 40(1), 27–33. http://doi.org/10.1016/j.eswa.2012.05.070.
Wang, W., Wang, H., & Song, Y. (2016). Ranking product aspects through sentiment analysis of online reviews. Journal of Experimental & Theoretical Artificial Intelligence, 1-20.
Wiegand, M., Klakow, D. (2012). Generalization Methods for In-Domain and Cross-Domain Opinion Holder Extraction. In: Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (Eacl´12) (pp. 325–335).
Wilson T, Wiebe J, Hoffmann P (2009) Recognizing contextual polarity: an exploration of features for phrase-level sentiment analysis. Computational Linguistics 35:399-433. doi: 10.1162/coli.08-012-R1-06-90.
Winkler S, Schaller S, Dorfer V, et al (2015) Data-based prediction of sentiments using heterogeneous model ensembles. Soft Computing 19:3401-3412. doi: 10.1007/s00500-014-1325-6.
Xie, S., Wang, G., Lin, S., & Yu, P. S. (2012). Review spam detection via temporal pattern discovery. In Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 823–831). http://doi.org/10.1145/2339530.2339662.
Yu Y, Wang X (2015) World Cup 2014 in the Twitter world: a big data analysis of sentiments in U.S. sports fans' tweets. Computers in Human Behavior 48:392-400. doi: 10.1016/j.chb.2015.01.075.
Yu, L. C., Wu, J. L., Chang, P. C., & Chu, H. S. (2013). Using a contextual entropy model to expand emotion words and their intensity for the sentiment classification of stock market news. Knowledge-Based Systems, 41(April), 89–97. http://doi.org/10.1016/j.knosys.2013.01.001.

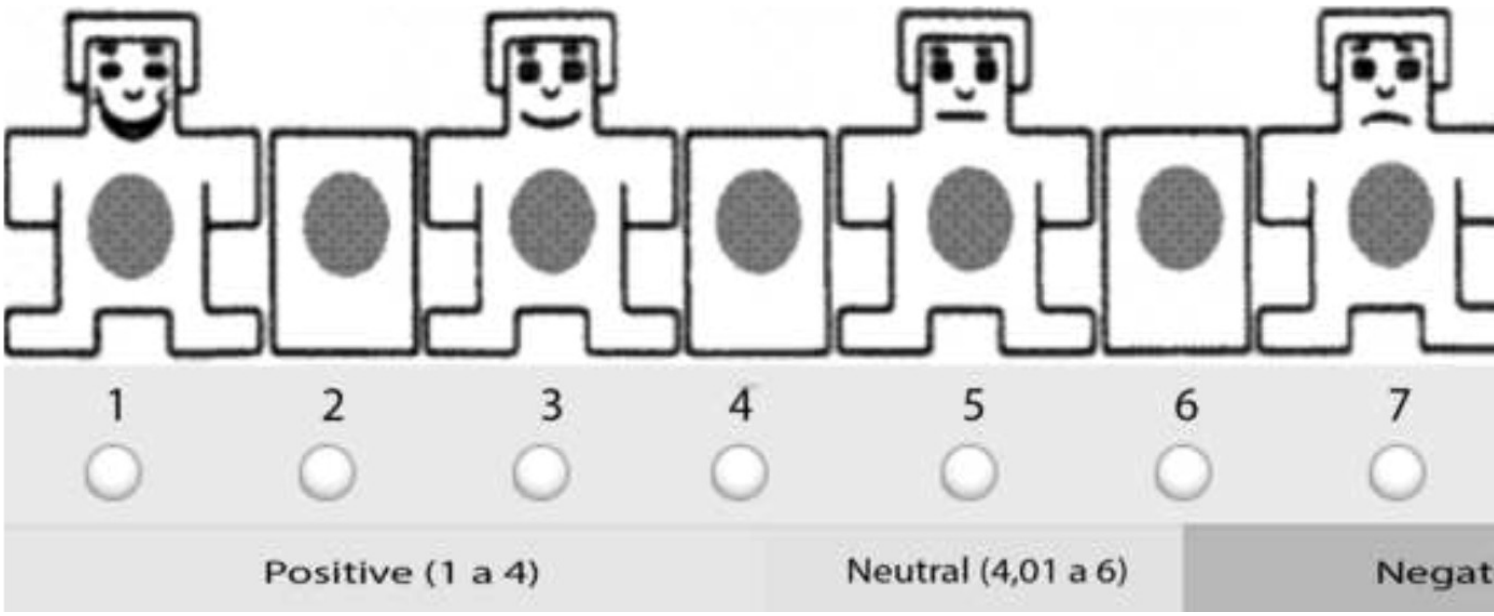
Compliance with Ethical Standard:

Funding: This study was funded by Coordination of Improvement of Higher Education, CAPES-Brazil (grant number BEX 2230/15-1), the Andalusian Excellence Projects (grant number P10-SEJ-6768) and the Spanish National Project (grant number TIN2013-40658-P).

Conflict of Interest: The author Steiner-Correa, A.F. declare that he has no conflict of interest. The author Viedma-del-Jesus, M.I. declare that she has no conflict of interest. The author López-Herrera, A.G. declare that he has no conflict of interest.

Ethical approval: This article does not contain any studies with human participants performed by any of the authors.

Figure



Table

Dataset	Language	Domain	Description	Class	Group	Courtesy of
2000Entities	English	Movie, Restaurant, Television, Politics, Sports, etc.	8,507 tweets collected from a total of about 2000 different entities from 20 different domains	Positive, negative, objective-not-spam and objective-spam	Training	Mukherjee et al. 20 http://www.mpi-in
Health Care Reform (HCR)	English	Health care reform	2,516 tweets containing the hashtag "#hcr" (health care reform)	Positive, negative, irrelevant or other	Training / Test	Saif et al. 2012; Sp https://bitbucket.or
Movies - UMICH SI650	English	Movies	40,138 tweets compiled by the University of Michigan for a SA tasks	Positive and negative	Training / Test	University of Mich https://inclass.kagg
Obama-McCain Debate (OMD)	English	Politics	3,238 tweets extracted from U.S. presidential TV debate in September 2008	Positive, negative and mix or unknown	Training	Shamma et al. 200 https://bitbucket.or
Stanford Twitter Dataset	English	No domain	1,600,000 tweets based on emoticons	Positive and negative	Training / Test	Go et al. 2009a http://help.sentime
SemEval_2015 Task 11	English	No domain	9,000 figurative tweets annotated with sentiment scores-ranging from -5...+5	Positive, neutral and negative	Training / Test	Rosenthal et al. 20 http://alt.qcri.org/s
Annotated-US2012-Election-Tweets	English	Politics	2,042 tweets annotated by 400 native English speakers extracted from August and September 2012	Trust, fear, surprise, sadness, disgust, anger, anticipation and joy.	Training	Mohammad et al. 2 http://saifmohamm
DAI-Labor English Dataset	English	No domain	7,200 tweets based on emoticons, annotated by 3 different researchers	Positive, neutral and negative	Training	Narr et al. 2012 http://dainas.aot.tu
RedBull Twitter Sentiment Dataset (RSD)	English	Beverage / Energy Drink	100 tweets annotated by 152 students, extracted in May 2014, based on the hashtag #givesyouwings from the RedBull company	Positive, neutral and negative	Training / Test	Steiner et al. 2016 http://mortero.ugr. rsd/
Noticias Globo	Portuguese (Brazil)	Feeds about Brazil and world	661 short messages extracted from www.globo.com	Joy, disgust, fear, anger and sadness	Training	Dosciatti et al. 201 http://www.ppgia.p
Política	Portuguese (Brazil)	Politics	567 tweets annotated by 3 different researchers	Positive, neutral and negative	Training	Nascimento et al. 2 Available directly

Entertainment	Portuguese (Brazil)	Entertainment, art and culture	384 tweets annotated by 3 different researchers	Positive, neutral and negative	Training	Nascimento et al. 2 Available directly
DAI-Labor Portuguese Dataset	Portuguese	No domain	1,800 tweets based on emoticons, annotated by 3 different researchers	Positive, neutral and negative	Training	Narr et al. 2012 http://dainas.aot.tu
General-TASS	Spanish	Politics, economics, communication and culture	68,000 tweets extracted from November 2011 until March 2012	6 classes: (P+) (P) (NEU) (N) (N+) and (NONE)	Training / Test	Román et al. 2015 http://www.sngular
Social-TV-TASS	Spanish	Sport	2,773 tweets during the Spanish Copa del Rey Football Final between Real Madrid and F.C. Barcelona on 16 April 2014	Positive, neutral and negative	Training / Test	Román et al. 2015 http://www.sngular
STOMPOL-TASS	Spanish	Politics	1,284 tweets extracted on the 23 and 24 April 2014	Positive, neutral and negative	Training / Test	Román et al. 2015 http://www.sngular
SpanishCorpus3100	Mexican Spanish	No domain	3,100 tweets annotated by 6 different researchers	5 classes (P+) (P) (NEU) (N) (N+)	Training	Baca-Gomez et al. Available directly
Modern Standard Arabic (MSA)	Arabic and Jordanian dialect	Politics and art	2,000 tweets annotated by 3 different researchers	2 Classes: Positive and Negative	Training	Assiri et al. 2015 https://archive.ics.u c+Sentiment+Anal
German Sentiment Dataset	German	German singers and musicians	500 short messages annotated by 3 native German speakers	Scores range -3...+3	Training	Momtazi 2012 www.hpi.uni- potsdam.de/fileadm N2/GermanSentim
DAI-Labor German Dataset	German	No domain	1,800 tweets based on emoticons, annotated by 3 different researchers	Positive, neutral and negative	Training	Narr et al. 2012 http://dainas.aot.tu
TWNews	Italian	Politics	3,228 ironic tweets between the 16 October 2011 and 3 February 2012, annotated by 3 researchers	5 classes: POS, NEG, HUM, MIXED and NONE	Training	Bosco et al. 2015 Available directly
TWSpino	Italian	Politics	1,159 ironic tweets extracted between July 2009 and February 2012, annotated by 3 researchers	5 classes: POS, NEG, HUM, MIXED and NONE	Training	Bosco et al. 2015 Available directly
Sentipolc Task - Evalita 2014	Italian	Politics / Generic	7,410 ironic tweets annotated by 3 researchers	Subjectivity, positive overall polarity, negative overall polarity, irony, positive literal polarity, negative literal polarity	Training	V. Basile et al. 201 http://www.di.unit

DAI-Labor French Dataset	French	No domain	1,797 tweets based on emoticons, annotated by 3 different researchers	Positive, neutral and negative	Training	Narr et al. 2012 http://dainas.aot.tu