Procesamiento del Lenguaje Natural, Revista nº 55, septiembre de 2015, pp 135-142 recibido 27-03-2015 revisado 23-04-2015 aceptado 20-05-2015

Is This Tweet Satirical? A Computational Approach for Satire Detection in Spanish*

¿Es satírico este tweet? Un método automático para la identificación del lenguaje satírico en español

Francesco Barbieri
Universitat Pompeu Fabra
francesco.barbieri@upf.edu

Francesco Ronzano
Universitat Pompeu Fabra
francesco.ronzano@upf.edu

Horacio Saggion
Universitat Pompeu Fabra
horacio.saggion@upf.edu

Resumen: La lingüística computacional está cada vez mas interesada en el procesamiento del lenguaje figurado. En este artículo estudiamos la detección de noticias satíricas en español y más específicamente la detección de sátira en mensajes de Twitter. Nuestro modelo computacional se basa en la representación de cada mensaje con un conjunto de rasgos diseñados para detectar el estilo satírico y no el contenido. Nuestros experimentos muestran que nuestro modelo siempre funciona mejor que un modelo de bolsa de palabras. También mostramos que el sistema es capaz de detectar este tipo de lenguaje independientemente de la cuenta de Twitter que lo origina.
Palabras clave: Detección Automática Sátira, Lenguaje Figurado, Análisis de Sentimientos

Abstract: Computational approaches to analyze figurative language are attracting a growing interest in Computational Linguistics. In this paper, we study the characterization of Twitter messages in Spanish that advertise satirical news. We present and evaluate a system able to classify tweets as satirical or not. To this purpose, we concentrate on the tweets published by several satirical and non-satirical Twitter accounts. We model the text of each tweet by a set of linguistically motivated features that aim at capturing the style more than the content of the message. Our experiments demonstrate that our model outperforms a word-based baseline. We also demonstrate that our system models global features of satirical language by showing that it is able to detect if a tweet contains or not satirical contents independently from the account that generated the tweet.
Keywords: Satire Detection, Figurative Language, Sentiment Analysis

1 Introduction

Computational approaches to analyze figurative language are attracting a growing interest in Computational Linguistics. Characterizing the figurative meaning of a sentence or text excerpt is extremely difficult to achieve by automated approaches. Properly dealing with figurative language constitutes a core issue in several research fields, including Human-Computer Interaction and Sentiment Analysis (Turney, 2002; Pang and Lee, 2008; Pak and Paroubek, 2010). Both of them would benefit of systems able to recognize figurative language. In the case of Sentiment Analysis for example, the literal sense of a text can be different and is often the opposite of its figurative meaning.

In this research we consider the case of satire, an important form of figurative language. Satire is a phenomena where humor and irony are employed to criticize and ridicule someone or something. Even if often misunderstood, "in itself, satire is not a comic device —it is a critique — but it uses comedic devices such as parody, exaggeration, slapstick, etc. to get its laughs." (Colletta, 2009). We focus on the study of satirical news in Spanish presenting a system able to separate satirical from non-satirical news. More specifically, we concentrate on Twitter

* The research described in this paper is partially funded by the SKATER-UPF-TALN project (TIN2012-38584-C06-03).

ISSN 1135-5948

messages published by several satirical and non-satirical Twitter accounts. As satirical Twitter accounts we consider "El Mundo Today" and "El Jueves", and as non-satirical ones the popular newspapers "El Mundo" and "El País".
Two examples respectively of satirical and non-satirical tweets are:

- **Satire** (from @elmundotoday)
  Ferran Adrià confiesa que su secreto es echarle a todo vinagre de Módena caramelizado.
  (Ferran Adrià confesses that his secret is adding to each dish caramelised Modena vinegar)
- **Non-Satire** (from @El_Pais)
  La enciclopedia de Ferran Adrià se pone en marcha. Más de 80 personas trabajarán en el nuevo proyecto del chef
  (The Ferran Adrià's Encyclopedia project begins. More than 80 people are going to work on the new chef's project).

As we read in the non-satirical tweet from the newspaper "El Pais", the popular Spanish chef Ferran Adrià is going to compile an Encyclopaedia of all the Spanish traditional dishes. The satirical news makes fun of this, saying that the only secret to make a good dish is adding Modena vinegar.

In this paper, we model each tweet by linguistically motivated features, which aim at capturing not the content but the style of the message. Our experiments demonstrate that our model outperforms a word-based baseline, in detecting if a tweet is satirical or not. We also show that our system detects satire independently from the Twitter account generating the messages.

The paper is organized as follows. The second Section is an overview of the state of the art on the characterization of satire. In Section 3 we describe the tools we used to process Spanish tweets. In Section 4 we introduce the features of our model. In Section 5 we describe the experiments we carried out to evaluate our model and present their results. In Section 6 we discuss the performance of our model. In the last section we present our conclusions and our plans for future work.

2 Related Work

Satire is a form of communication where humor and irony are used to criticize someone's behavior and ridicule it. Satirical authors may be aggressive and offensive, but they "always have a deeper meaning and a social signification beyond that of the humor"(Colletta, 2009). Satire loses its significance when the audience do not understand the real intents hidden in the ironic dimension. Indeed, the key message of a satirical utterance lays in the figurative interpretation of the ironic sentence. Satire has been often studied in literature (Peter, 1956; Mann, 1973; Knight, 2004; LaMarre, Landreville, and Beam, 2009), but rarely with a computational approach. The work of Burfoot and Baldwin (2009) attempts to computationally model satire in English. They retrieved news-wires documents and satiric news articles from the web, and build a model able to recognize satirical articles. Their approach included standard text classification (Binary feature weights and Bi-normal separation feature scaling), lexical features (including profanity and slang) and semantic validity. To characterize the semantic validity of an excerpt, they identify its named entities and query the web for the conjunction of those entities, expecting that satirical conjunctions were less frequent than the ones from non-satirical news.

As said above, irony plays a key role in satire. The standard definition of irony is "saying the opposite of what you mean" (Quintilien and B., 1953). Grice (1975) believes that irony is a rhetorical figure that violates the maxim of quality, while Giora (1995) says that irony can be any form of negation with no negation markers. Wilson (2002) defined irony as echoic utterance that shows a negative aspect of someone's else opinion. Utsumi (2000) and Veale (2010a) stated that irony is a form of pretence that is violated. Since 2010 researchers designed models to detect irony automatically. Veale (2010b) proposed an algorithm for separating ironic from non-ironic similes in English, detecting common terms used in this ironic comparison. Reyes et. al (2012; 2013) proposed a model to detect irony and humour in English tweets, pointing out that Ambiguity (computed using perplexity on Google n-gram) and skip-grams which capture word sequences that contain (or skip over) arbitrary gaps, are the most informative features. Barbieri and Saggion (2014) designed an irony detection system that avoided the use of the

word-based features. However, irony has not been studied intensively in languages other than English. Few studies addressed irony detection in non-English languages like Portuguese (Carvalho et al., 2009; De Freitas et al., 2014), Dutch (Liebrecht, Kunneman, and van den Bosch, 2013) and Italian (Barbieri, Ronzano, and Saggion, 2014).

3 Data and Text Processing

We parse the textual contents of tweets in order to extract relevant linguistic and semantic features as described in this Section. We use the tool Freeling (Carreras et al., 2004) to perform sentence splitting, tokenization, POS tagging, and Word Sense Disambiguation (WSD) of tweets. WSD in Freeling relies on the Spanish Wordnet distributed by the TALP Research Centre. The Spanish Wordnet is mapped by means of the Inter-Lingual-Index to the English Wordnet 3.0 whose synset IDs are in turn characterized by sentiment scores by means of SentiWordnet[1]. In order to define the usage frequency of the words of a tweet, we use a corpus we built from a dump of the Spanish Wikipedia[2] as of May 2014.

In order to train and test our system we retrieved tweets from four twitter accounts (two satirical and two non-satirical) from June 2014 to January 2014. We gathered tweets from the satirical accounts "El Mundo Today" and "El Jueves". The non-satirical tweets were retrieved from the Twitter accounts of real newspapers: "El Mundo" and "El Pais". For each account we gathered 2,766 tweets, hence the final corpus includes 11,064 tweets. After downloading the tweets we filtered them by removing tweets that were not relevant to our study (for instance: "Buy our new issue" or "Watch the video"). We left only tweets that advertize actual news (satirical or non-satirical). We share this dataset[3] as a list of tweet IDs since per Twitter policy it is not possible to share tweets contents.

4 Our Method

This Section describes the two systems we compare with respect to their ability to classify the tweets of our dataset as satirical or not. The first system (Section 4.1) is the actual satire-detection system we present in this paper; it relies on lexical and semantic features to characterize each word of a tweet. The second system (Section 4.2) constitutes our baseline to evaluate our real approach and model tweets by relying on lemma occurrences (BOW approach). Both systems exploit Support Vector Machine[4] (Platt and others, 1999) to classify tweets as satirical or not.

4.1 Satire Detection Model

We implement a similar model to (Barbieri and Saggion, 2014) for irony detection. We characterize each tweet by seven classes of features:Frequency, Ambiguity, Part Of Speech, Synonyms, Sentiments, Characters, and Slang Words These features aim to describe intrinsic aspects of the words included in satiric tweets. The interesting propriety of the intrinsic word features is that they do not rely on words-patterns hence detect more abstract (and Twitter account-independent) traits of satire.

4.1.1 Frequency

We access the frequency corpus (see Section 3) to retrieve the frequency of each word of a tweet. Thus, we derive three types of Frequency features: *rarest word frequency* (frequency of the most rare word included in the tweet), *frequency mean* (the arithmetic average of all the frequency of the words in the tweet) and *frequency gap* (the difference between the two previous features). These features are computed including all the words of each tweet. We also calculate these features by considering only Nouns, Verbs, Adjectives, and Adverbs.

4.1.2 Ambiguity

To model the ambiguity of the words in the tweets we use the WordNet Spanish synsets associated to each word. Our hypothesis is that if a word has many meanings (synsets associated) it is more likely to be used in an ambiguous way. For each tweet we calculate the *maximum number of synsets* associated to a single word, the *mean synset number* of all the words, and the *synset gap* that is the difference between the two previous features. We determine the value of these features by including all the words of a tweet as well as

[1] http://sentiwordnet.isti.cnr.it/

[2] We thank Daniel Ferrés for his help.

[3] http://sempub.taln.upf.edu/tw/sepln2015/

[4] We relied on the LibLINEAR implementation, http://www.csie.ntu.edu.tw/ cjlin/liblinear

by considering only Nouns, Verbs, Adjectives or Adverbs.

4.1.3 Part Of Speech

The features included in the Part Of Speech (POS) group are designed to capture the syntactic structure of the tweets. The features of this group are eight and each one of them counts the number of occurrences of words characterized by a certain POS. The eight POSs considered are Verbs, Nouns, Adjectives, Adverbs, Interjections, Determiners, Pronouns, and Apposition.

4.1.4 Synonyms

We consider the frequencies (for each language its own frequency corpora, see Section 3) of the synonyms of each word in the tweet, as retrieved from WordNet. Then we compute, across all the words of the tweet: the *greatest* and the *lowest number of synonyms* with frequency higher than the one present in the tweet, the *mean number of synonyms* with frequency greater / lower than the frequency of the related word present in the tweet. We determine also the greatest / lowest number of synonyms and the mean number of synonyms of the words with frequency greater / lower than the one present in the the tweet. We compute the set of Synonyms features by considering both all words of the tweet together and only the words belonging to each one of the four POSs listed before.

4.1.5 Sentiments

The sentiment of the words in tweets is important for two reasons: to detect the *sentiment* (e.g. if tweets contain mainly positive or negative terms) and to capture unexpectedness created by a negative word in a positive context or vice versa. Relying on Sentiment lexicons (see Section 3) we computed the *number of positive / negative words*, the *sum of the intensities of the positive / negative scores of words*, the *mean of positive / negative score of words*, the *greatest positive / negative score*, the *gap between the greatest positive / negative score and the positive / negative mean*. Moreover we simply count (and measure the ratio of) the *words with polarity* not equal to zero, to detect subjectivity in the tweet. As previously done, we compute these features by considering both all the words of a tweet and only Nouns, Verbs, Adjectives, and Adverbs.

4.1.6 Characters

Even if Satirical news try to mimic the same punctuation style than non-satirical newspapers, we also wanted to capture the punctuation style of the authors, and the type of characters employed in a tweet. This is because punctuation is very important in social networks: ellipses can be sign of satire for instance, or a full stop of negative emotion. Each feature that is part of this set is the number of occurrences of a specific punctuation mark, including: ".", "!", "?", "$", "%", "&", "+", "-", "=". We also compute the numbers of Uppercase and Lowercase characters, and the length of the tweet.

4.1.7 Bad Words

Since Twitter messages often include *slang words*, we count them as they may be used often in satirical and informal messages (we compiled a list of 443 "slang words" in Spanish).

4.2 Word-Based Baseline

All the features belonging to this group are useful to model common word-patterns. These features are used to train our baseline system that classifies tweets as satirical or not, in order to carry out a comparative evaluation with our actual system that relies on the other groups of features described in the previous section. We compute the five word-based features: *lemma* (lemmas of the tweet), *bigrams* (combination of two lemmas in a sequence) and *skip 1/2/3 gram*. For each of these feature we keep the 1,000 most frequent occurrences in each training set considered (we carry out several experiments considering distinct training sets, thus considering distinct feature occurrence in each experiment, see Section 5).

5 Experiments and Results

In order to test the performances of our system we run two kind of balanced binary classificati on experiments where the two classes are "satire" and "non-satire". Our dataset includes two newspaper accounts, N1 and N2, and two satirical news accounts, S1 and S2. In the **first binary balanced classification experiment**, we train the system on a dataset composed of 80% of tweets from one of the newspaper accounts and 80% of tweets from one of the satirical accounts (5,444 tweets in total). Then we test the system on a dataset that includes 20% of the

Train	Test	Word Based	Intrinsic Features	All Features
N1 vs S1	N2 vs S2	0.622	**0.754**	0.727
N1 vs S2	N2 vs S1	0.563	0.712	**0.723**
N2 vs S1	N1 vs S2	0.592	**0.805**	0.709
N2 vs S2	N1 vs S1	0.570	**0.778**	0.737
N1-N2 vs S1-S2	N1-N2 vs S1-S2	0.735	0.814	**0.852**

Table 1: F1 of each newspaper/satirical account combination, where N1="El Pais", N2="El Mundo", S1="El Mundo Today", and S2="El Jueves". In **bold** the best results (not by chance confirmed by two-matched-samples t-test with unknown variances) between word-based and Intrinsic features.

tweets of a newspaper account that is different from the one used in the training and 20% of the tweets of a satirical account that has not been used for training. The final size of our testing set is 1,089 tweets. We run the following configurations:

- Train: N1 (80%) and S1 (80%)
  Test: N2 (20%) and S2 (20%)
- Train: N1 (80%) and S2 (80%)
  Test: N2 (20%) and S1 (20%)
- Train: N2 (80%) and S1 (80%)
  Test: N1 (20%) and S2 (20%)
- Train: N2 (80%) and S2 (80%)
  Test: N1 (20%) and S1 (20%)

With these configurations we never use tweets from the same account in both the training and testing datasets, thus we can evaluate the ability of our system to detect satire independently from the features of a specific Twitter account. As a consequence we avoid the *account modeling / recognition* effect, as the system is never trained on the same accounts where it is tested. Moreover, in order to study the learning progression in relation to the number of tweets, we divide each training set in ten folds and test the systems using 1 to 10 folds to train it. In other words, we start using a tenth of the training tweets, and progressively add a tenth of tweets more until reaching the size of the whole training set.

In the **second binary balanced classification experiment**, the training set is composed of 80% of the tweets of each account. The test includes the remaining 20% of the tweets of each account. Hence the training set includes 8,710 tweets and the test set includes 2,177 tweets.

For the two classification experiments just introduced, we test three models: the baseline (BoW, see Section 4.2), our model (Section 4.1), and the union of them. The results are reported in Table 1. The reader can note that in each experiment our system outperforms the baseline. In the first configuration the baseline achieves F1 between 0.563 to 0.622 with a mean of 0.586 in the four combinations. In the same experiment our system obtains better F1 in every configuration, with values in the range 0.712-0.805 with a mean of 0.762. In Table 2 we show the results of the three systems when only a portion of the training set is used (refer to the first column, Fold). For each fold and each system we report the mean of the four account combinations. We can see that even if the BoW slightly improves its performance when adding more tweets to the training set, our system always performs better. Additionally, our system achieves high accuracy even when using a tenth of the training tweets: with only 544 tweets the F1 of our system (the mean of the four combinations) is 0.743.

Fold	BoW	Intr.	All
1	0.526	**0.743**	0.729
2	0.530	**0.754**	0.709
3	0.556	**0.755**	0.713
4	0.559	**0.762**	0.725
5	0.565	**0.759**	0.729
6	0.579	**0.755**	0.726
7	0.571	**0.756**	0.728
8	0.576	**0.760**	0.722
9	0.576	**0.757**	0.721
10	0.586	**0.762**	0.724

Table 2: First binary classification experiments with progressive training set size (from 544 (fold 1) to 5440 (fold 10) tweets. For each experiment is reported the mean of the F1 of the four account combinations.

In the second experiment (the union of all the accounts) the baseline model improves its

performance, but our model is still better. The F1 are respectively 0.735 for the baseline model and 0.814 for our model.

In order to understand the contribution of each feature in the two models we computed the information gain on the training set of the second binary classification experiment (where tweets from all the accounts are included). The best 20 features of each model are shown in Table 3 and Table 4. The most relevant words to detect satire are slang expressions, articles and specific nouns; in the list verbs are not present and "serio" ("serious") is the only adjective. The best features of our model are the ones of the Character group, followed by Part of Speech, Frequency, Slang Words and Ambiguity. To notice, Synonyms and Sentiment groups do not contribute as much as the other groups.

IG	Lemma Feat.	Translation
0.023	manda_güevos	(slang)
0.011	en	in
0.011	de	of
0.011	que	that
0.009	saludos	greetings
0.007	por	to
0.007	sobre	on
0.007	le	him/her
0.006	el_minuto	the_minute
0.006	partido	match
0.006	uno	one
0.006	el_por	the_to
0.006	serio	serious
0.006	porque	because
0.005	méxico	mexico
0.005	mucho_gracia	thanks_a lot
0.005	te	you
0.005	tú	you
0.005	el_sobre	the_envelope

Table 3: Best lemma-based features ranked computing the information gain on the N1-N2 vs S1-S2 training set.

6 Discussion

Our system outperforms the baseline in each experiment settings. In Table 1 we can see that in the cross-account experiments (training on two accounts and testing in the other two accounts) the baseline is not able to recognize satire. Indeed lemma-based features are useful to model the vocabulary of a specific Twitter account instead of abstracting less domain/account dependent features. This is also proven by the high results obtained by the baseline in the second experiment (last row of Table 1), where both the training and test sets include tweets from the same accounts (all the accounts considered). The difference between the mean score on the first configuration (first four rows of Table 1) and the second one (last row of Table 1) is 0.15. On the other hand our model is more stable and obtains good results in each configuration. The difference in performance is lower (0.52) suggesting that our model does not depend on specific accounts. However in the first experiment our model does not obtain the same results in all the combinations: it is hard to explain exactly why, more experiments are needed. With a closer look, we can see that the best configuration is obtained when training on "El Mundo vs El Mundo Today" and testing on "El Pais vs El Jueves" (inverting these train and test datasets we obtain the worse configuration). This suggests that the combination "El Mundo vs El Mundo Today" includes diverse examples than "El Pais vs El Jueves". Indeed it is possible to detect satire with high accuracy when training on the first dataset but not in the second one. Vice versa, a system trained on "El Pais vs El Jueves" recognizes fewer ironic tweets of "El Mundo vs El Mundo Today".

Another important point is how fast our system learns to detect satirical tweets. Our system achieves very good performances if trained on a dataset made of only 544 tweets, and with less than half of the tweets available for training (Table 2, fold 4, 2176 tweets) our system obtains the best F1.

The information gain experiments give interesting information on how the systems work. The baseline model that relies on lemma-based features depends on accounts and topics. Indeed, the lemma-based feature with most information gain is a Spanish slang ("manda_güevos") as in the account "El Jueves" is often used while never present in the newspaper accounts. The other relevant features of the baseline model are nouns that depend on topics like "el partido" ("the match") and "méxico". In our model (Table 4) the most important features are the ones relative on the style of the message (Character and Frequency features). Indeed, features like the length of the message, the case and the length of the words, and number of exclamation points have high information gain. The structure of the message (Part of

IG	Feature Group	Feature Name
0.231	Characters	Length
0.087	Characters	Uppercase
0.080	Characters	First uppercase word
0.078	Characters	Lowercase
0.072	Part of Speech	Number Nouns
0.040	Characters	Longest word
0.038	Characters	Mean word length
0.037	Characters	Number of !
0.034	Part of Speech	Number Apposition
0.026	Frequency	Frequency gap (Nouns)
0.025	Frequency	Rarest Adjective
0.024	Ambiguity	Mean number Synsets
0.023	Part of Speech	Number Numbers
0.022	Frequency	Frequency Mean (Nouns)
0.021	Part of Speech	Number Pronouns
0.020	Frequency	Rarest Noun
0.017	Badwords	Badwords Ratio
0.017	Characters	Number of -
0.017	Frequency	Frequency mean (Adjectives)
0.015	Frequency	Frequency gap
0.013	Ambiguity	Max Number Synsets

Table 4: Best 20 features considering the Information Gain calculated on the N1-N2 vs S1-S2 training set (second experiment configuration where all the accounts are included).

Speech group) is also important as features like number of nouns and apposition are relevant on satire detection. The ambiguity feature plays an important role too, and the satirical tweets present words with greater polisemy (number of synsets associated) than newspaper tweets. Finally, a simple but relevant feature is the presence of "slang words", than obviously are more used in the satirical news.

We were not able to compare our approach with other satire detection systems (Burfoot and Baldwin, 2009) since approaches and dataset are very different. An important incompatibility is we only used lexical information, while Burfoot and Baldwin also included meta-information by searching the web. The other relevant difference was the dataset: they considered whole satirical and non-satirical articles, while we only use messages at most 140 characters long (tweets). Moreover, their research was on English articles.

7 Conclusion and Future Work

In this paper we present a system for the automatic detection of Spanish satirical news. We retrieve text from Twitter accounts of newspapers and satirical Twitter accounts. Our system classifies tweets by relying on linguistically motivated features that aim at capturing not the content but the style of the message. We show with cross-account experiments (experiments that never share tweets of the same Twitter accounts among training and test sets) that our system detects satire with good accuracy considerably improving performance with respect to a Bag of Words baseline. Bag of Words baselines are able to model the dictionary of specific accounts more than to detect satire.

In the future we aim to improve our model adding new features (e.g. distributional semantic) and increase our dataset by incorporating new Twitter accounts so as to perform a more extensive evaluation of our results.

References

Barbieri, F., F. Ronzano, and H. Saggion. 2014. Italian Irony Detection in Twitter: a First Approach. In *The First Italian Conference on Computational Linguistics CLiC-it 2014*, page 28.

Barbieri, F. and H. Saggion. 2014. Modelling Irony in Twitter. In *Student Research Workshop at EACL*, pages 56–64, Gothenburg, Sweden, April. ACL.

Burfoot, C. and T. Baldwin. 2009. Automatic satire detection: Are you having a laugh? In *Proceedings of the*

ACL-IJCNLP 2009 conference short papers, pages 161–164. ACL.

Carreras, Xavier, Isaac Chao, Lluis Padró, and Muntsa Padró. 2004. Freeling: An open-source suite of language analyzers. In *Language Resources and Evaluation Conference*.

Carvalho, P., L. Sarmento, M. Silva, and E. de Oliveira. 2009. Clues for detecting irony in user-generated contents: oh...!! it's so easy;-). In *Proceedings of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion*, pages 53–56. ACM.

Colletta, L. 2009. Political satire and postmodern irony in the age of stephen colbert and jon stewart. *The Journal of Popular Culture*, 42(5):856–874.

De Freitas, L. A, A. Vanin, D. Hogetop, M. Bochernitsan, and R. Vieira. 2014. Pathways for irony detection in tweets. In *Proceedings of the 29th Annual ACM Symposium on Applied Computing*, pages 628–633. ACM.

Giora, R. 1995. On irony and negation. *Discourse processes*, 19(2):239–264.

Grice, H. 1975. Logic and conversation. *1975*, pages 41–58.

Knight, C. 2004. *The literature of satire*. Cambridge University Press.

LaMarre, H., K. Landreville, and M. Beam. 2009. The irony of satire political ideology and the motivation to see what you want to see in the colbert report. *The International Journal of Press/Politics*, 14(2):212–231.

Liebrecht, Christine, Florian Kunneman, and Antal van den Bosch. 2013. The perfect solution for detecting sarcasm in tweets# not. *WASSA 2013*, page 29.

Mann, J. 1973. *Chaucer and Medieval Estates Satire: The Literature of Social Classes and the General Prologue to the Canterbury Tales*. Cambridge University Press Cambridge.

Pak, A. and P. Paroubek. 2010. Twitter as a corpus for sentiment analysis and opinion mining. In *Language Resources and Evaluation Conference*.

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

Peter, J. 1956. Complaint and satire in early english literature.

Platt, John et al. 1999. Fast training of support vector machines using sequential minimal optimization. *Advances in kernel methodssupport vector learning*, 3.

Quintilien and Harold E. B. 1953. *The Institutio Oratoria of Quintilian. With an English Translation by HE Butler*. W. Heinemann.

Reyes, A., P. Rosso, and T. Veale. 2013. A multidimensional approach for detecting irony in Twitter. *Language Resources and Evaluation*, pages 1–30.

Reyes, Antonio, Paolo Rosso, and Davide Buscaldi. 2012. From humor recognition to irony detection: The figurative language of social media. *Data & Knowledge Engineering*, 74:1–12.

Turney, P. 2002. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th ACL*, pages 417–424.

Utsumi, Akira. 2000. Verbal irony as implicit display of ironic environment: Distinguishing ironic utterances from nonirony. *Journal of Pragmatics*, 32(12):1777–1806.

Veale, T. and Y. Hao. 2010a. An ironic fist in a velvet glove: Creative misrepresentation in the construction of ironic similes. *Minds and Machines*, 20(4):635–650.

Veale, T. and Y. Hao. 2010b. Detecting Ironic Intent in Creative Comparisons. In *ECAI*, volume 215, pages 765–770.

Wilson, Deirdre and Dan Sperber. 2002. Relevance theory. *Handbook of pragmatics*.