Procesamiento del Lenguaje Natural, Revista nº 70, marzo de 2023, pp. 85-98 recibido 14-12-2022 revisado 03-02-2023 aceptado 13-02-2023

# When humour hurts: linguistic features to foster explainability

## *Cuando el humor duele: características linguísticas para ganar en explicabilidad*

**Lucía I. Merlo[1], Berta Chulvi[1,2], Reynier Ortega[1], Paolo Rosso[1]**

[1]Universitat Politècnica de València, Spain
[2]Universitat de València, Spain
lumer1@inf.upv.es, berta.chulvi@upv.es, prosso@dsic.upv.es, rortega@prhlt.upv.es

**Abstract:** The main objective of this research is to use different features for the textual representation of humorous texts and detect which are the characteristics that distinguish non-offensive jokes from the highly offensive ones. For this purpose, we use the data from the HaHackaton task in which jokes are annotated according to their degree of offensiveness. A new classification task is created by using two subsets of the jokes: the non-offensive ones and the highly offensive ones.The features with statistically significant differences in the two groups are used. By applying an ablation test, the most relevant features are used for a second classification task, showing that it is possible to obtain the same results with fewer computational resources.
**Keywords:** Humour, offensive language, computational lingüistics.

**Resumen:** El objetivo de esta investigación es utilizar distintas características para representar los textos humorísticos y detectar cuáles son las que mejor distinguen los chistes no ofensivos de los muy ofensivos. Se utiliza los datos de la tarea HaHackaton en la que los chistes están anotados según su grado de ofensa. Se diseña un nuevo problema de clasificación con dos conjuntos de chistes: los nada ofensivos y los muy ofensivos. Los clasificadores se entrenaron con las características que presentan diferencias significativas en las dos clases. Mediante la aplicación de un *ablation test* se identificaron las más relevantes que se han utilizado en una segunda tarea de clasificación mostrando que es posible obtener los mismos resultados con menos recursos computacionales.
**Palabras clave:** Humor, lenguaje ofensivo, lingüística computacional.

## *1 Introduction*

When a society begins to overcome its prejudices, humour is one of the spaces in which these prejudices remain longer. As the social psychologist Michael Billig stands in his research about humour: if the collective laughter has a shameful, darker side, then, there is a lot that we may wish to hide from ourselves (Billig, 2005). This argument builds upon the insights of Bergson (Bergson, 1900) and Freud (Freud, 1960) who suggest that humour -and mainly the part of humour which serves to ridicule- ensures that members of society routinely comply with the customs and habits of their social milieu to avoid being the objects of jokes.

Certainly, humour has many facets and multiple effects on a social and personal life (Martin and Ford, 2018). It can be rebellious, kicking against the dictates of social norms and defending minority identities (Dobai and Hopkins, 2020). Also, it is well known that humour has an important beneficial function for personal life (Ripoll and Casado, 2010). But sometimes, certain type of humour is more than a simple joke. It has important consequences for some minority groups at personal and at societal level (Ford et al., 2008). For this reason, this research aims to detect how is the language that conveys offense in humour as a first step towards understanding how humour is an effective language device by means of prejudice and stereotypes can be maintained and perpetuated.

The prejudice norm theory (Ford and Ferguson, 2004) stands that *disparagement humour* function as a source of self-regulation for people high in prejudice because it creates

ISSN 1135-5948 DOI 10.26342/2023-70-7 

a normative climate of tolerance of discrimination. This could be the reason why offense towards certain groups is well canalised through humour. The effects of these offensive jokes spill over into other spaces with far more serious consequences. For example, research about sexism has demonstrated that for men high in hostile sexism, sexist humour can have important social consequences, for example on rape proclivity (Romero-Sánchez et al., 2017).

Humour as a way to offend is not limited to intergroups relation, but it is also used at a interpersonal level. This kind of humour has been defined as *adversarial humour* (Veale, Feyaerts, and Brône, 2006). This humour occurs when up to a certain point, jokes have the underlying goal of weaken the opponent's position in a given social interaction.

Jokes are also a cultural product very sensitive to the passage of time. Something amusing twenty years ago, nowadays might be considered boring, aggressive or even hateful. Detecting offense in humour is a complex matter (Merlo, 2022). A joke can have abusive language but not being hurtful and the opposite: it can be hurtful without being explicitly abusive (Yin and Zubiaga, 2022). If detecting when a joke is hurtful is complex it is even more difficult to explain the results obtained in a classification task on the basis of the linguistic characteristics of the texts.

### 1.1 Objectives and research questions

Nowadays, social media platforms are widely extended all over the world and are often used to express hate speech camouflaged into jokes, trying to hide underlying negative attitudes. For hate speech monitoring activities, it is crucial to distinguish between offensive and non-offensive humour. This distinction is also relevant when analysing the communicative climate in a given community. We are conscious that, deep learning models achieve very impressive results in many NLP tasks in terms of effectiveness (Grover and Goel, 2021; Song et al., 2021; Potamias, Siolas, and Stafylopatis, 2020; González, Hurtado, and Pla, 2020), but often they may be quite complex from an explainability point of view. Then, our objective is to identify linguistic patterns present in hurtful humour that could help to automatically recognise these types of communication. This characterization of the language used in offensive humour could serve to gain on explainability when deep learning models are applied in humour recognition tasks. With this aim this work addresses four research questions:

- RQ1. Which are the features that distinguish non-offensive humour from the offensive ones?
- RQ2. How do three standard machine learning classifiers perform using these linguistic features?
- RQ3. Which of these linguistics features contribute more to the classification task?
- RQ4. Is it possible to obtain similar or better results employing only the most relevant linguistic features?

## *2 Related work*

One of the first researches on humour recognition considering linguistic features was presented by Mihalcea and Strappavara (Mihalcea and Strapparava, 2005). The authors carried out a study for distinguishing humorous and non-humorous texts, using a computational approach for humour recognition. Furthermore, humorous examples consisted in one-liners while non-humorous texts were extracted from three resources: Reuters news headlines, proverbs and texts from British National Corpus (BNC). In English context, one-liner is an idiom to refer a short joke or witty remarks. Through classification systems, it was possible for them to detect which linguistic features were relevant. Specifically, classifiers were trained with stylistic features (alliterations, antonyms and slangs), content features and a combination of both. The results showed that stylistic markers help to distinguish a large number of one-liners jokes from Reuters news headlines and from BNC's texts, but not from proverbs. The authors suggest that content features help to differentiate jokes and proverbs although their stylistic similarity, but do not help to distinguish jokes from Reuters news headlines and BNC's texts. They remark on how humorous data mainly include words that refer to human scenarios (man, woman, I, you, person) and negative forms of words (isn't, doesn't, bad).

Sjöbergh and Araki tried to determine whether a text is a joke without considering the meaning of it (Sjöbergh and Araki,

2007). They used a corpus of 6,100 one-liner jokes and phrases from the British National Corpus for non-humorous examples. The features considered were text similarity (word overlap between the training instances and the text to classify, applying a novel weighting scheme), most common words within jokes (e.g. animals are particularly frequent), measure of ambiguity in a phrase, stylistic features (rimes, repeated words, use of you/I/he/she, negations) and idiomatic expressions (e.g. It's a piece of cake). The obtained results yielded that common words in jokes seemed to be the most useful feature for humour distinction, whereas stylistic features did not seem to provide a substantial contribution. Despite this, the article discusses how humorous texts differ from others without recognizing the meaning, although the features extracted from content markers were considered as highly relevant.

A weakness of the research mentioned so far, is that the humorous and the non-humorous texts come from quite different sources, e.g. one-liners vs sentences from British National Corpus. These sources present significant differences between them, regarding topic, vocabulary and target audience. Trying to overcome this issue (Reyes et al., 2010) studied a corpus of online comments retrieved from the Slashdot news website. The authors used a selection of 600,000 comments annotated by users into four categories: funny, informative, insightful and negative. The classification models were trained with linguistic features related to sexual content, semantic ambiguity, polarity, emotions, slang and emojis. By computing a multi-label classification, the authors examined which of the features contributed the most to humour recognition. They observed that the distinction between funny and informative categories was more challenging than the differentiation between funny and insightful and funny and negative ones. Regarding the features, slangs terms and emojis helped to improve humour recognition.

On the other hand, tasks related to humour recognition have attracted many researchers to the field. In English we can find the HashtagWars task in SemEval-2017 (Potash, Romanov, and Rumshisky, 2017) and in SemEval-2020 (Hossain et al., 2020) related with humor in headlines. In SemeEval-2021, the HaHackathon task (Meaney et al., 2021) proposed to distinguish between humorous and non-humorous texts while including several subtasks. The second task mentioned, also set as a subtask the prediction of the rate of offense in texts as we explain in detail in Section 3. In Spanish we find the HAHA task in 2018 (Castro, Chiruzzo, and Rosa, 2018), in 2019 (Chiruzzo et al., 2019) and in 2021 (Chiruzzo et al., 2021). All these tasks proposed a principal task of humour recognition and different subtasks. Furthermore, in the 2021 edition of HAHA, the organisers proposed to predict a funniness score value for each tweet, the mechanism by which the tweet conveys humour belongs to a set of classes (irony, wordplay, hyperbole, or shock) and the content of which the joke is based on, with the main target related to racist jokes, sexist jokes, dark humor, dirty jokes, among others until fifteen categories.

In these evaluation forums, the top-ranking teams made extensive use of pre-trained language models such as BERT, ERNIE (Zhang et al., 2019), ALBERT (Lan et al., 2020) or RoBERTa (Zhuang et al., 2021). These approaches had an excellent performance in accuracy. Still, they cannot distil linguistic knowledge valuable for understanding how language devices (particularly humour) convey offensiveness, stereotypes and prejudice. As a consequence, we do not have an overly recent knowledge about which linguistic features are the most important ones to distinguish offensive humour from non-offensive humour. Moreover, recent works (Ortega-Bueno, Rosso, and Medina Pagola, 2022; Frenda et al., 2022; Cignarella et al., 2020) have shown that reinforcing the deep learning models with linguistic knowledge helps to improve their overall performance. As a result, the aim of the following experiments is to provide some insights in this direction.

## *3 Data and preprocessing*

### 3.1 Data

For this research we used the dataset of the *HaHackaton* task from *HaHackaton, Detecting and Rating Humor and Offense* organised at SemEval-2021 (Meaney et al., 2021). In the original dataset 80% of texts are originated in Twitter and unsettled 20% is obtained from the Kaggle *Short Jokes* dataset (Moudgil, 2017). Some keywords referring of-

fense to certain groups were used in the data collection strategy. Complete examples of offensive keywords and jokes with them can be found in appendix A in Tables 11 and 12. A total of 10,000 texts compose the original dataset of the *HaHackaton* task. Text annotation was done by US citizenship participants belonging to the following age groups: 18-25, 26-40, 41-55, 56-70. Each text was annotated by 5 members of each group. The task organisers instructed annotators to indicate if the text has the intention to be humorous in a 1 to 5 scale. As a second question they asked if the text was generally offensive in a scale of 1 to 5. They instructed annotators to consider as generally offensive a text which targets a person or group of people, simply for belonging to a certain group or a text that a large number of people were likely to be offended by. The offense rating of each text is the average of all ratings given by the annotators, including 'no offense' as 0.

Our research makes use of "offense rating" annotation to create a new classification task into two new categories: the non-offensive humour vs the most offensive humour. In order to create these two new datasets, we used the offense rating of each joke in the original dataset (0-5) and we created four groups of texts, each one corresponding to a quartile of the offense score variable. For our analysis and for the classification experiments, only the outermost groups are used. Therefore, our dataset is composed by the first quartile (the non-offensive jokes) and the fourth quartile (the highly offensive jokes) of the original dataset of the *HaHackaton* task. Specifically, the non-offensive set has 1,601 instances and the highly offensive set is composed by 1,504 examples. To answer the first three research questions in Section 4 and in Section 5 we use only the training set (1,253 non-offensive and 1,210 highly offensive instances) of the *HaHackaton* dataset. We keep the test set of *HaHackaton* (348 non-offensive and 294 highly offensive) to answer RQ4 in Section 7.

### 3.2 Checking the manual annotation

We applied several exploratory strategies to evaluate the quality of the manual annotation of the dataset. Firstly, the Spearman correlation between offense rating and humour rating scores over humorous data has been calculated in the two sets of jokes. With a $\rho$ of -0.27 and a p-value $< .001$, we observe that the annotators tend to consider a text with greater amounts of humour if the level of offense in it is low or null.

As a second strategy to evaluate the quality of annotations regarding offense rating variable, we proceed in two steps. The first one consists in computing features from several linguistic resources, for instance: *SenticNet* (Cambria et al., 2016), *Textblob* (Loria and et.al, 2020), *SentiWordNet* (Baccianella, Esuli, and Sebastiani, 2010), *VADER* (Hutto and Gilbert, 2014), *ANEW* (Warriner, Kuperman, and Brysbaert, 2013) and *AFINN* (Nielsen, 2011). The second step consists of calculating either the Mann-Whitney U test or the Wilcoxon Signed-Ranked test, attending whether the observations are paired or not, over the quantitative features, taking as independent variable the offense group (non-offensive jokes vs highly offensive jokes). The complete results can be found in Appendix A in Tables 13 and 14. In summary, it can be seen that we find statistical differences between the two classes of humour in sentiment score, values, polarity, abusive language and subjectivity using the above-mentioned resources.

### 3.3 Text representation

Linguistic feature extraction conforms the core of this analysis. Hence, vectorized representation of features are achieved it through the *Stanza tool* (Peng Qi and Manning., 2020) and lexicons: *Binary Lexicon of abusive words* (Wiegand et al., 2018), *Hurtlex* (Bassignana, Basile, and Patti, 2018), *EmoSenticNet* (Bandyopadhyay et al., 2013), *SentiSense* (de Albornoz, Plaza, and Gervás, 2012) and *LIWC* (Tausczik and Pennebaker, 2010).

To extract part-of-speech tags, syntactic & morphological information, the *Stanza tagger* for English is used. Each term is assigned to a tag (noun, pronoun, adjective, tenses, 1st/2nd/3rd persons). The information regarding punctuation symbols is also computed by the *Stanza tagger*, by applying it over the original texts.

Variables related to affective and content information are constructed from lexical resources. The feature extraction procedure is equal for both of them. Tokens within tweets, are compared to the list of terms contained

in each one of the lexical resources used. Afterwards, we computed the number of times each word of the terms-list appear within the document. The *LIWC* resource also enables to extract syntactic & morphological markers, besides the affective and the content ones. Finally, the features are obtained by dividing the frequency of terms found in the tweet over the tweet length in terms of number of words. As a result, texts are represented as a frequency weighted term vector. Hence, each $i$-value of the linguistic feature corresponds to the rate of occurrence of determined category inside the $i$-tweet.

## *4 On offensive humour attributes*

### 4.1 Statistical analysis

To select the most suitable features to represent the texts in the experimental phase we decided to identify the ones in which the offense label (non-offensive vs highly offensive) introduces statistically significant differences between the distributions of quantitative data.

Firstly, the Spearman correlation has been computed in order to determine whether or not values of the same feature from each class are independent. If the null hypothesis is true, observations are not paired and the Mann-Whitney U test is used. Rejecting the null hypothesis means that observations are paired and the Wilcoxon Signed-Ranked test is computed. This analysis was carried out by considering a p-value with a significance of 0.05.

The features included in the next section are those with a statistically significant difference with a confidence level of 95%, between the two groups of non-offensive jokes and offensive ones.

### 4.2 Features for offense

With the selected features, a classification of these into three groups has been carried out, distinguishing affective, content and syntactic & morphological markers.

**Syntactic & morphological markers** reflect the style of writing and the types of terms used. These are elements which provide of coherence within texts (Weth, 2020) by relating terms within a sentence. In addition, part-of-speech markers such as nouns, adjectives, adverbs, verbs, auxiliary verbs, persons and tenses are considered as part of these markers. Results are shown in Table 1.

**Affective markers** covers sentiments, emotions and attitude terms within a sentence. In this case, the features derived from sentiment markers quantify negative and positive words/terms, according to the mentioned lexical resources. A similar procedure is followed for features associated with personal states and emotions such as anger, disgust, joy, like, love, sadness, surprise. Results are included in Table 2.

**Content markers** indicate terms related to the content of a sentence: words from diverse categories used in LIWC dictionary (social, biology or religion) and hateful words, negative stereotypes and moral defects from Hurtlex dictionary, among other categories (see Table 3).

As observed in Table 1, among syntactic & morphological features, first personal pronouns, both singular and plural, and second personal pronouns in singular have a higher ratio of occurrence in non-offensive jokes than in offensive ones. However, the third personal pronoun in plural follows an opposite pattern. Although being highly present in offensive and non-offensive tweets, variables regarding articles (a, an, the), adjectives (cruel, bored, awful) and auxiliary verbs (am, has, might), have a higher frequency in offensive texts. Uniquely considering these variables, articles have the most outstanding difference of occurrence between both types of texts, being mostly used in offensive contexts. As articles define a noun as specific or unspecific, their appliance in line with the explanation about the use of personal pronouns, it might be useful to increment the distance between the sender and the object of the joke. For instance: *"You **the** bomb." "No, you **the** bomb." In America, **a** compliment. In **the** Middle East, **an** argument.* Adjectives also have a wider presence in offensive texts. By taking into account this context, and the fact that these words make reference to an attribute of a thing/person, terms used tend to be hurtful, like in this example: *What **do** you get if you cross **an illiterate african american** with **an illegal hispanic** immigrant looking for **a** green card? **A** United States soldier.*

When inspecting the results for affective features (see Table 2) we observe that negative emotions (anger, disgust, fear and sad-

| | | | Non-offensive | | Highly-offensive | |
|---|---|---|---|---|---|---|
| **Lexicon** | **Feature** | **p-value** | **Mean** | **Variance** | **Mean** | **Variance** |
| LIWC | I | 1.35E-45 | 0.0706 | 0.0051 | 0.0351 | 0.0036 |
| LIWC | Personal Pronouns | 5.50E-11 | 0.1268 | 0.0062 | 0.0964 | 0.0061 |
| LIWC | Article | 9.58E-10 | 0.0748 | 0.0038 | 0.0915 | 0.0047 |
| PoS | Adjective | 1.87E-07 | 0.0816 | 0.004 | 0.0968 | 0.0049 |
| LIWC | They | 2.76E-07 | 0.0064 | 0.0004 | 0.0127 | 0.001 |
| LIWC | Prepositions | 3.87E-07 | 0.1037 | 0.0043 | 0.0893 | 0.0035 |
| LIWC | Auxiliary Verb | 2.67E-06 | 0.0902 | 0.003 | 0.1007 | 0.0031 |
| PoS | 1st Plural Person | 3.25E-06 | 0.0033 | 0.0002 | 0.0012 | 0.0001 |
| PoS | Adverbs | 2.91E-05 | 0.0566 | 0.0033 | 0.048 | 0.0031 |
| PoS | Noun | 8.87E-05 | 0.2511 | 0.0088 | 0.2379 | 0.0092 |
| PoS | 2nd Person Singular | 4.93E-03 | 0.0013 | 0.0001 | 0.0005 | 3.0e-05 |

Table 1: Syntactic & morphological features belonging to non- and highly offensive jokes.

| | | | Non-offensive | | Highly-offensive | |
|---|---|---|---|---|---|---|
| **Lexicon** | **Feature** | **p-value** | **Mean** | **Variance** | **Mean** | **Variance** |
| EmoSenticNet | Surprise | 2.14E-13 | 0.0409 | 0.0032 | 0.0639 | 0.0057 |
| SentiSense | Fear | 2.31E-11 | 0.0078 | 0.0005 | 0.015 | 0.001 |
| LIWC | Positive Emotions | 2.38E-10 | 0.0322 | 0.0021 | 0.0223 | 0.0015 |
| LIWC | Inhibition | 0.00014 | 0.0056 | 0.0003 | 0.0036 | 0.0002 |
| LIWC | Anxiety | 1.65E-04 | 0.0039 | 0.0002 | 0.0027 | 0.0002 |
| LIWC | Affective Processes | 2.70E-04 | 0.0583 | 0.004 | 0.0487 | 0.0031 |
| SentiSense | Disgust | 3.00E-04 | 0.03 | 0.002 | 0.0366 | 0.0022 |
| LIWC | Anger | 1.67E-03 | 0.0087 | 0.0005 | 0.0127 | 0.0009 |
| SentiSense | Sadness | 1.08E-02 | 0.0033 | 0.0002 | 0.0053 | 0.0004 |
| SentiSense | Like | 1.80E-02 | 0.0276 | 0.0016 | 0.0244 | 0.0015 |
| SentiSense | Joy | 1.80E-02 | 0.0058 | 0.0003 | 0.0094 | 0.0006 |
| SentiSense | Love | 2.74E-02 | 0.0061 | 0.0003 | 0.0044 | 0.0003 |

Table 2: Affective features belonging to non- and highly offensive jokes.

ness) appear to be highly present through offensive jokes, in contrast to non-offensive ones. Moreover, the offensive set presents a higher amount of terms related to surprise, an emotion that could be either positive or negative. Additionally, affective processes from LIWC and positive emotions in general tend to appear mostly in non-offensive jokes.

A different trend is visible for terms associated to the emotion of joy. Results expose a greater occurrence in offensive texts than in non-offensive ones. When inspecting the linguistic resource the joy variable is extracted from, it is observed that the *gay* term is associated with this emotion −as it was in old English−, although it also is nowadays a term associated to a sexual orientation, as shown in the following examples: *I am **laughing** at these ladies waking up and being like Hey wanna become **gay icons** today?* and *Why do we hate making up **gay jokes**? Because it's always a pain in the as**.

Results regarding the content features are observed in Table 3. Content features are related to the topic of the jokes. It is noticeable that words associated to biology, humans, sexual, social, religion, negative stereotypes, moral and behavioural defects, swear words and ethnic slurs are mostly used in highly offensive jokes than in non-offensive ones. A good example of this kind of use is the following: *Where do most **black people** work? In **jail**.*

The most notorious differences between non-offensive and offensive texts are observed in features with jokes regarding sexuality (gay, lesbian, prostitute), religion (Jewish, christian, Christmas), swear words, negative stereotypes and ethnic slurs (Mexican, Chinese, black people) and moral or behavioural defects (jail, death).

| | | | Non-offensive | | Highly-offensive | |
|---|---|---|---|---|---|---|
| **Lexicon** | **Feature** | **p-value** | **Mean** | **Variance** | **Mean** | **Variance** |
| LIWC | Social | 1.38E-12 | 0.1161 | 0.0088 | 0.1418 | 0.0091 |
| LIWC | Biology | 1.09E-14 | 0.0363 | 0.0031 | 0.0534 | 0.0038 |
| LIWC | Quantifiers | 1.38E-07 | 0.0206 | 0.0011 | 0.0304 | 0.0019 |
| LIWC | Humans | 2.11E-38 | 0.0103 | 0.0006 | 0.0283 | 0.0017 |
| LIWC | Sexual | 2.39E-38 | 0.0038 | 0.0002 | 0.0198 | 0.0016 |
| LIWC | See | 1.23E-09 | 0.0111 | 0.0008 | 0.0197 | 0.0015 |
| LIWC | Exclusive | 1.88E-08 | 0.0213 | 0.0012 | 0.0143 | 0.0009 |
| LIWC | Leisure | 2.63E-07 | 0.0201 | 0.0015 | 0.0136 | 0.001 |
| LIWC | Religion | 5.86E-15 | 0.0026 | 0.0002 | 0.0115 | 0.0012 |
| Hurtlex | Negative stereotypes and ethnic-slurs | 8.64E-40 | 0.0004 | 2.2e-05 | 0.0105 | 0.0008 |
| Hurtlex | Moral & behavioural defects | 2.56E-23 | 0.0023 | 0.0001 | 0.01 | 0.0006 |
| LIWC | Swear words | 6.95E-27 | 0.0009 | 5.0e-05 | 0.0082 | 0.0006 |

Table 3: Content features belonging to non- and highly offensive jokes.

## *5 Classification experiments*

This section focuses on the classification of the jokes as non-offensives or highly offensives. The execution of experiments is performed by dividing the training set of the dataset in 80% for training and 20% for testing. As it is a binary classification, the offensive set is considered as the positive class, and the non-offensive set as the negative class. The classifiers applied are: Support Vector Machine (SVM), Random Forests (RF) and Logistic Regression (LR). For evaluating the performance of the classifiers, measures of accuracy, precision (PR) and $F_1$-score were computed with a five-fold cross validation. As baselines we employed SVM, RF and RL with Bag of Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF) for text representation. Their results are shown in Tables 4 and 5. In a second set of experiments, the three classifiers were trained with the 35 most relevant linguistic features identified as statistically significant in the exploratory analysis (see Tables 1, 2 and 3). Table 6 provides the results obtained by each classifier trained with the most relevant linguistic features and Table 8 shows the rates achieved.

Compared to the results obtained with the baselines, the three classifiers perform better using the features proposed in this study, especially when classifying offensive samples. As we can see in Table 5, for the baselines the main problem is that highly offensive jokes are classified as non-offensive. In relation to the precision metric, it is observed that the best performance is obtained by RF when using all the linguistic features, similarly when using BoW and TF-IDF.

## *6 Ablation test*

In order to quantify the contribution provided by each group of features in the classifiers performance, an ablation test has been done. Table 7 shows that, in general, all classifiers perform worse when removing any group of features. However, content features are the most important for the classification task. The removal of this set of features decreases in a substantial manner the $F_1$-score for all classifiers. This can be observed in FPR, FNR, TPR and TNR showed in Table 8. These values worsened when removing this set of features: increasing the FPR and decreasing the TNR. As a consequence, the recall decreases and the precision is altered.

Removing syntactic & morphological features also generates a drop in $F_1$-score metric, similar for all the classifiers (see Table 7) but not as strong as for the content features case. The performance of the classifiers differ when removing this group of features. All models present a greater percentage of offensive instances misclassified (increase in the FNR) and less capability of classify positive cases properly (decrease in the TPR), as shown in Table 8. However, FPR and TNR improve in SVM and LR, contrary to RF which does not present any variation in these metrics.

Regarding the affective features, a non ex-

| | Non-offensive | Highly-offensive | | |
|---|---|---|---|---|
| **Model** | **F1-score** | **F1-score** | **F1-macro** | **Accuracy** |
| SVM | 0.70 | 0.61 | 0.66 | 0.66 |
| RF | 0.66 | 0.05 | 0.35 | 0.49 |
| LR | 0.70 | 0.61 | 0.66 | 0.66 |

Table 4: F1-score & Accuracy of the baselines with BoW+TF*IDF.

| | **FPR** | **FNR** | **TPR** | **TNR** | **PR** |
|---|---|---|---|---|---|
| SVM | 0.18 | 0.48 | 0.52 | 0.82 | 0.75 |
| RF | 0.004 | 0.98 | 0.02 | 0.99 | 0.86 |
| LR | 0.18 | 0.48 | 0.52 | 0.82 | 0.75 |

Table 5: Classification of the baselines with BoW+TF*IDF.

| | Non-offensive | Highly-offensive | | |
|---|---|---|---|---|
| **Model** | **F1-score** | **F1-score** | **F1-macro** | **Accuracy** |
| SVM | 0.76 | 0.72 | 0.74 | 0.74 |
| RF | 0.77 | 0.75 | 0.76 | 0.76 |
| LR | 0.74 | 0.72 | 0.73 | 0.73 |

Table 6: F1-score & Accuracy with the 35 linguistic features proposed.

| | **SVM** | **RF** | **LR** |
|---|---|---|---|
| | **F1-macro** | **F1-macro** | **F1-macro** |
| All features | 0.74 | 0.76 | 0.73 |
| Affective | 0.73 (↓ 0.01) | 0.72 (↓ 0.04) | 0.75 (↑ 0.02) |
| Syntactic & morphological | 0.72 (↓ 0.02) | 0.73 (↓ 0.03) | 0.72 (↓ 0.01) |
| **Content** | **0.65 (↓ 0.09)** | **0.66 (↓ 0.1)** | **0.66 (↓ 0.07)** |

Table 7: Ablation test for SVM, RF and LR.

| | | **FPR** | **FNR** | **TPR** | **TNR** | **PR** |
|---|---|---|---|---|---|---|
| **All features** | SVM | 0.17 | 0.34 | 0.66 | 0.83 | 0.81 |
| | RF | 0.16 | 0.31 | 0.69 | 0.84 | 0.82 |
| | LR | 0.22 | 0.32 | 0.68 | 0.78 | 0.77 |
| **Affective** | SVM | 0.16 | 0.36 | 0.64 | 0.84 | 0.81 |
| | RF | 0.19 | 0.36 | 0.64 | 0.81 | 0.78 |
| | LR | 0.18 | 0.31 | 0.69 | 0.82 | 0.80 |
| **Syntactic** & | SVM | 0.15 | 0.39 | 0.61 | 0.85 | 0.81 |
| **morphological** | RF | 0.16 | 0.37 | 0.63 | 0.84 | 0.80 |
| | LR | 0.20 | 0.36 | 0.64 | 0.80 | 0.78 |
| **Content** | **SVM** | **0.34** | **0.37** | **0.63** | **0.66** | **0.66** |
| | **RF** | **0.33** | **0.35** | **0.65** | **0.67** | **0.68** |
| | **LR** | **0.33** | **0.35** | **0.65** | **0.66** | **0.67** |

Table 8: Classification metrics for the classifiers with all features and for the ablation test.

pected result is observed for the LR model (Table 7). This classifier obtains better results when removing this set. Looking at FPR and TNR (Table 8), there is a slight improvement in their values. However, it is noticeable how the decrease of misclassified instances in the positive class (lower FPR) widely contributes to the increase of 0.03 in the precision score in comparison with the LR model trained with all features (Table 8). This res-

ult must be explored in more detail as a future work. A first hypothesis could be related to the confusion effect that the emotion *joy* could introduce in the sense we explained in Section 4.2. One reason could be that, in this situation, SVM and RF models are more robust than the LR model.

## 7 Classification with less features

The results of the ablation test show that content features were the most relevant for the classification task. Then, a second set of experiments has been carried out only with the content features to answer the RQ4. In these experiments, the classifiers were trained with the complete training set and the test set was used to evaluate their performance. The results of these experiments are provided in Tables 9 and 10.

When the classifiers use the content features RF obtains the same results for accuracy and $F_1$-score as when using all of the linguistic features, while SVM and LR increase their performance. Taking into account that LR and SVM with linear kernel as hyperparameter are both linear classifiers, the results observed for SVM and LR when trained with all features could be due to the effect that multicollinearity (Bayman and Dexter, 2021) has over both models. That is to say, their vulnerability towards small changes in the data and difficulties on identify feature importance. To explore if the correlation between features could justify the effect of multicollinearity, an exploratory analysis has been done. We could see that a content feature like *social* from LIWC presents a significant correlation ($rho = 0.42$) with a syntactic & morphological feature as it is *personal pronouns*, and content feature *moral & behavioural defects* from Hurtlex correlates with *disgust* ($rho = 0.24$) and with *fear* ($rho = 0.35$) from SentiSense. Therefore, some of these relations could be introducing redundant information, and worsening the classifiers performance.

## 8 Discussion of results

Regarding RQ1, we identified in our preliminary analysis which are the features that distinguish non-offensive humour from the offensive one. As we see in Tables 1, 2 and 3, a set of content, syntactic & morphological and affective features are useful to differentiate between the two classes of humour. Among the content features *negative stereotypes*, *moral and behavioural defects* and *swear words* are used in a very different way in both classes of humour. A possible reason for this result, could be that offensive humour is mainly reserved to ridicule minority groups or people that present certain behaviours that contradict mainstream values.

Among the syntactic & morphological features, we observe that the first person pronouns, both singular and plural, and second person pronouns in singular have a higher ratio of occurrence in non-offensive jokes than in offensive ones. A possible explanation for this result can rely on the depersonalization of the sender when saying something hurtful. This can be used as a mechanism to take off responsibilities of conveying offensive jokes and removes any possible feeling of guilty. However, third person pronoun in plural follows an opposite pattern, being more frequent in offensive jokes. This result allows to think that offensive jokes share linguistic patterns with other communicative phenomena related to prejudice, as hate speech (Chulvi, Toselli, and Rosso, 2022) and extremism (Torregrosa et al., 2022) where a more frequent use of "they" narratives has been observed. Regarding the features that capture aspects related to emotions, we observe that negative ones (anger, disgust, fear and sadness) appear to be highly present through offensive jokes in comparison to the non-offensive ones. Therefore, at least in this dataset, we can conclude that the offensive jokes are used to convey negative emotions towards particular groups, values and behaviours.

In response to RQ2, we observe that all the classifiers perform better using the proposed linguistic features in comparison with the baselines and all of them perform better distinguishing non-offensive humour from the offensive one. We used standard machine learning classifiers avoiding transformers, given that our main objective focuses on a descriptive analysis of the features that could contribute to the explainability of the results.

In this sense, a result from the ablation tests is that content features are the ones that contribute in a substantial manner for all classifiers (RQ3). This role of content features is in line with some first researches in this area, that showed the importance of cer-

| | | FPR | FNR | TPR | TNR | PR |
|---|---|---|---|---|---|---|
| **BoW+TF-IDF** | SVM | 0.14 | 0.49 | 0.51 | 0.86 | 0.76 |
| | RF | 0.01 | 0.90 | 0.10 | 0.99 | 0.86 |
| | LR | 0.14 | 0.49 | 0.51 | 0.86 | 0.76 |
| **Content** | SVM | 0.23 | 0.26 | 0.74 | 0.77 | 0.73 |
| | RF | 0.20 | 0.28 | 0.72 | 0.80 | 0.75 |
| | LR | 0.24 | 0.25 | 0.75 | 0.76 | 0.72 |

Table 9: Classification metrics of the baselines and the classifiers with content features.

| | | Non-offensive | Highly-offensive | | |
|---|---|---|---|---|---|
| | **Model** | **F1-score** | **F1-score** | **F1-macro** | **Accuracy** |
| **BoW+TF-IDF** | SVM | 0.76 | 0.61 | 0.68 | 0.70 |
| | RF | 0.72 | 0.19 | 0.45 | 0.58 |
| | LR | 0.76 | 0.61 | 0.69 | 0.70 |
| **Content** | SVM | 0.77 | 0.74 | 0.75 | 0.75 |
| | RF | 0.78 | 0.73 | 0.76 | 0.76 |
| | LR | 0.77 | 0.74 | 0.75 | 0.75 |

Table 10: F1-score & Accuracy of the classifiers with the baselines and with content features.

tain words in the detection of humour (Mihalcea and Strapparava, 2005) even when the strategy was the opposite (Sjöbergh and Araki, 2007).

Regarding to RQ4, we may conclude that it is possible to adopt a strategy with less computational resources, as long as a previous study is carried out, as shown in Section 4 and in the ablation test (Section 6). It is relevant to consider that the set of the most relevant features, the ones that we called content features in our experiments, come from two different linguistic resources: LIWC and Hurtlex.

## *9 Conclusions and future work*

In this work we have represented two sets of jokes (non-offensive and highly offensive ones) with the use of computational linguistics resources such as LIWC, Hurtlex, SentiSense and EmoSentiNet. The goal was to identify which linguistic features are used differently in non-offensive and offensive humour. We have used these features in a classification task. Subsequently, by applying an ablation test, we were able to detect which groups of features contribute the most. We have used these features in a strategy for using less computational resources, showing that it is possible to obtain the same performance. From a social science point of view, these results allows us to take a step towards a research program that explore how offensive humour is used to construct otherness and underpin prejudice.

As future work, we plan to compare our results with Transformers-based models, although instead of comparing the effectiveness, we plan to focus on identifying similarities and differences between the features highlighted by the attention mechanism and our linguistic features. Moreover, we plan to integrate the most relevant linguistic features in Transformers and deep learning-based models to help explainability during their decision-making process when detecting hurtful humour.

## *Acknowledgements*

This work was done in the framework of the FairTransNLP research project (PID2021-124361OB-C31) funded by MCIN/AEI/10.13039/501100011033 and by ERDF, EU A way of making Europe. It has been developed with the support of valgrAI - Valencian Graduate School and Research Network of Artificial Intelligence and the Generalitat Valenciana, and co-funded by the European Union.

## *References*

Baccianella, S., A. Esuli, and F. Sebastiani. 2010. SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Seventh International Conference*

*on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Bandyopadhyay, S., D. Das, N. Howard, A. Hussain, A. Gelbukh, and S. Poria. 2013. Enhanced SenticNet with affective labels for concept-based opinion mining. *IEEE Intelligent Systems*, 28(02):31–38.

Bassignana, E., V. Basile, and V. Patti. 2018. Hurtlex: A Multilingual Lexicon of Words to Hurt. In CEUR-WS, editor, *Proceedings of the Fifth Italian Conference on Computational Linguistics CLiC-it 2018*, volume 2253, pages 1–6, Turin, Italy.

Bayman, E. O. and F. Dexter. 2021. Multicollinearity in logistic regression models. *Anesthesia & Analgesia*, 133(2):362–365. https://journals.lww.com/anesthesia-analgesia/Fulltext/2021/08000/Multicollinearity_in_Logistic_Regression_Models.12.aspx.

Bergson, H. 1900. *Le rire:essai sur la signification du comique*. Félix Alcan, Paris, France.

Billig, M. 2005. *Laughter and Ridicule: toward a social critique of humour*. Sage, London.

Cambria, E., S. Poria, R. Bajpai, and B. Schuller. 2016. SenticNet 4: A semantic resource for sentiment analysis based on conceptual primitives. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2666–2677, Osaka, Japan.

Castro, S., L. Chiruzzo, and A. Rosa. 2018. Overview of the HAHA Task: Humor analysis based on human annotation at IberEval 2018. In *IberEval@ SEPLN*, pages 187–194.

Chiruzzo, L., S. Castro, M. Etcheverry, D. Garat, J. Prada, and A. Rosa. 2019. Overview of HAHA at IberLEF 2019: Humor analysis based on human annotation. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019)*, CEUR Workshop Proceedings. CEUR-WS.

Chiruzzo, L., S. Castro, S. Góngora, A. Rosa, J. A. Meaney, and R. Mihalcea. 2021. Overview of HAHA at IberLEF 2021: Detecting, Rating and Analyzing Humor in Spanish. *Procesamiento del Lenguaje Natural*, 67(0):257–268.

Chulvi, B., A. H. Toselli, and P. Rosso. 2022. Fake news and Hate Speech: Language in Common. Technical report, I International Seminar on Artificial Intelligence and disinformation. https://arxiv.org/pdf/2212.02352.pdf.

Cignarella, A. T., V. Basile, M. Sanguinetti, C. Bosco, P. Rosso, and F. Benamara. 2020. Multilingual irony detection with dependency syntax and neural models. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1346–1358, Barcelona, Spain (Online), December. International Committee on Computational Linguistics.

de Albornoz, J. C., L. Plaza, and P. Gervás. 2012. SentiSense: An easily scalable concept-based affective lexicon for sentiment analysis. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3562–3567, Istanbul, Turkey. European Language Resources Association (ELRA).

Dobai, A. and N. Hopkins. 2020. Humour is serious: Minority group members' use of humour in their encounters with majority group members. *European Journal of Social Psychology*, 50(2):448–462.

Ford, T. and M. Ferguson. 2004. Social Consequences of Disparagement Humor: A Prejudiced Norm Theory. *Personality and Social Psychology Review*, 8(1):79–94.

Ford, T. E., C. F. Boxer, J. Armstrong, M. Moya, and J. R. Edel. 2008. More than "just a joke": the prejudice-releasing function of sexist humor. *Personality & social psychology bulletin*, 34(2):159–70.

Frenda, S., A. T. Cignarella, V. Basile, C. Bosco, V. Patti, and P. Rosso. 2022. The unbearable hurtfulness of sarcasm. *Expert Systems with Applications*, 193:116398.

Freud, S. 1960. *Jokes and their Relation to the Unconscious*. Norton, Harmondsworth, England.

González, J. Á., L. F. Hurtado, and F. Pla. 2020. Transformer based contextualization of pre-trained word embeddings for irony detection in Twitter. *Information Processing and Management*, 57:1–15.

Grover, K. and T. Goel. 2021. Haha@ iberlef2021: Humor analysis using ensembles of simple transformers. In *IberLEF@ SEPLN*, pages 883–890.

Hossain, N., J. Krumm, M. Gamon, and H. Kautz. 2020. SemEval-2020 task 7: Assessing humor in edited news headlines. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 746–758, Barcelona (online). International Committee for Computational Linguistics.

Hutto, C. and E. Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media*, 8(1):216–225.

Lan, Z., M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*.

Loria, S. and et.al. 2020. Textblob: Simplified Text Processing. https://textblob.readthedocs.io/en/dev/. Online; accessed 01 March 2022.

Martin, R. A. and T. E. Ford. 2018. Chapter 1 - introduction to the psychology of humor. In R. A. Martin and T. E. Ford, editors, *The Psychology of Humor (Second Edition)*. Academic Press, second edition edition, pages 1–32.

Meaney, J., S. Wilson, L. Chiruzzo, A. Lopez, and W. Magdy. 2021. Semeval 2021 task 7: Hahackathon, detecting and rating humor and offense. In *SemEval 2021 Task 7: HaHackathon, Detecting and Rating Humor and Offense*, pages 105–119.

Merlo, L. I. 2022. When humour hurts: A computational linguistic approach. Final degree project. Technical report. http://hdl.handle.net/10251/188166.

Mihalcea, R. and C. Strapparava. 2005. Making computers laugh: Investigations in automatic humor recognition. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 531–538, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Moudgil, A. 2017. Short jokes. https://www.kaggle.com/datasets/abhinavmoudgil95/short-jokes. Online; accessed 01 March 2022.

Nielsen, F. Å. 2011. A new ANEW: evaluation of a word list for sentiment analysis in microblogs. In M. Rowe, M. Stankovic, A.-S. Dadzie, and M. Hardey, editors, *Proceedings of the ESWC2011 Workshop on "Making Sense of Microposts": Big things come in small packages*, volume 718 of *CEUR Workshop Proceedings*, pages 93–98.

Ortega-Bueno, R., P. Rosso, and J. E. Medina Pagola. 2022. Multi-view informed attention-based model for Irony and Satire detection in Spanish variants. *Knowledge-Based Systems*, 235:107597.

Peng Qi, Yuhao Zhang, Y. Z. J. B. and C. D. Manning. 2020. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In *In Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

Potamias, R. A., G. Siolas, and A. G. Stafylopatis. 2020. A transformer-based approach to irony and sarcasm detection. *Neural Computing and Applications*, 32(23):17309–17320.

Potash, P., A. Romanov, and A. Rumshisky. 2017. SemEval-2017 task 6: #HashtagWars: Learning a sense of humor. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 49–57, Vancouver, Canada. Association for Computational Linguistics.

Reyes, A., M. Potthast, P. Rosso, and B. Stein. 2010. Evaluating humour features on web comments. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation*

*(LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Ripoll, R. M. and I. Q. Casado. 2010. Laughter and positive therapies: Modern approach and practical use in medicine. *Revista de Psiquiatría y Salud Mental (English Edition)*, 3(1):27–34.

Romero-Sánchez, M., H. Carreterio-Dios, J. L. Megías, M. Moya, and T. Ford. 2017. Sexist Humor and Rape Proclivity: The Moderating Role of Joke Teller Gender and Severity of Sexual Assault. *Violence against women*, 23(8):951–972.

Sjöbergh, J. and K. Araki. 2007. Recognizing humor without recognizing meaning. In F. Masulli, S. Mitra, and G. Pasi, editors, *Applications of Fuzzy Sets Theory*, pages 469–476, Berlin, Heidelberg. Springer Berlin Heidelberg.

Song, B., C. Pan, S. Wang, and Z. Luo. 2021. DeepBlueAI at SemEval-2021 task 7: Detecting and rating humor and offense with stacking diverse language model-based methods. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 1130–1134, Online, August. Association for Computational Linguistics.

Tausczik, Y. and J. Pennebaker. 2010. The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology*, 29:24–54.

Torregrosa, J., G. Bello-Orgaz, E. Martínez-Cámara, J. D. Ser, and D. Camacho. 2022. A survey on extremism analysis using natural language processing: definitions, literature review, trends and challenges. *Journal of Ambient Intelligence and Humanized Computing*.

Veale, T., K. Feyaerts, and G. Brône. 2006. The cognitive mechanisms of adversarial humor. *Humor-international Journal of Humor Research - HUMOR*, 19:305–339.

Warriner, A. B., V. Kuperman, and M. Brysbaert. 2013. Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior Research Methods*, 45(4):1191–1207.

Weth, C. 2020. Distinguishing Syntactic Markers From Morphological Markers. A Cross-Linguistic Comparison. *Frontiers in Psychology*, 11:2082.

Wiegand, M., J. Ruppenhofer, A. Schmidt, and C. Greenberg. 2018. Inducing a lexicon of abusive words – a feature-based approach. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1046–1056, New Orleans, Louisiana. Association for Computational Linguistics.

Yin, W. and A. Zubiaga. 2022. Hidden behind the obvious: Misleading keywords and implicitly abusive language on social media. *Online Social Networks and Media*, 30:100210.

Zhang, Z., X. Han, Z. Liu, X. Jiang, M. Sun, and Q. Liu. 2019. ERNIE: Enhanced language representation with informative entities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1441–1451, Florence, Italy. Association for Computational Linguistics.

Zhuang, L., L. Wayne, S. Ya, and Z. Jun. 2021. A robustly optimized BERT pre-training approach with post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1218–1227, Huhhot, China. Chinese Information Processing Society of China.

## *A Appendix*

| Target | Keywords |
|---|---|
| Sexism | She, woman, mother, girl, b*tch, he, man, blond, p*ssy |
| Body | Fat, thin, tall, short, bald |
| Origin | Mexican, Mexico, Irish, Ireland, Chinese, Asian |
| Sexual orientation | Gay, lesbian, homo, LGBT, trans |
| Racism | Black, white people, nig** |
| Ideology | Feminism, lefty |
| Religion | Muslim, Jewish, Jew, Catholic, Jesus, Christmas |
| Health | Blind, deaf, r*tard, dyslexic, wheelchair |

Table 11: Offensive keywords in the *HaHackaton* dataset (Meaney et al., 2021).

| Target | Keyword = Target |
|---|---|
| A **fat** woman just served me at McDonalds and said "Sorry about the wait". I replied and said, "Don't worry, you'll lose it eventually". | Yes |
| Don't worry if a **fat** guy comes to kidnap you... I told Santa all I want for Christmas is you. | No |

Table 12: Examples of jokes with keywords mentioned in the HaHackaton overview paper (Meaney et al., 2021).

| Tool or Lexicon | Feature | p-value | Mean | Variance |
|---|---|---|---|---|
| SentiWordNet | Sentiment Score | 0.0021 | 0.5188 | 0.0061 |
| AFINN | Valence Score | 2.5e-11 | 0.6446 | 0.0041 |
| VADER | Sentiment score | 1.282e-11 | 0.0824 | 0.1908 |
| TextBlob | Polarity score | 1.31e-07 | 0.0708 | 0.0758 |
| | Subjectivity score | 4.83e-07 | 0.4114 | 0.0995 |
| ANEW | Valence score | 2.3e-10 | 5.7582 | 0.2126 |
| | Dominance score | 4.23e-10 | 5.5608 | 0.0968 |
| | Arousal score | 0.011 | 4.0808 | 0.1116 |
| Lexicon of abusive words extended | Score | 0.003 | 0.4715 | 0.013 |

Table 13: Tagger features in non-offensive tweets in the humorous subset.

| Tool or Lexicon | Feature | p-value | Mean | Variance |
|---|---|---|---|---|
| SentiWordNet | Sentiment Score | 0.0021 | 0.5084 | 0.0061 |
| AFINN | Valence Score | 2.5e-11 | 0.6236 | 0.0049 |
| VADER | Sentiment score | 1.282e-11 | -0.0476 | 0.1801 |
| TextBlob | Polarity score | 1.31e-07 | 0.0164 | 0.0621 |
| | Subjectivity score | 4.83e-07 | 0.35 | 0.0811 |
| ANEW | Valence score | 2.3e-10 | 5.6276 | 0.243 |
| | Dominance score | 4.23e-10 | 5.4778 | 0.1049 |
| | Arousal score | 0.011 | 4.1241 | 0.1504 |
| Lexicon of abusive words extended | Score | 0.003 | 0.4836 | 0.0175 |

Table 14: Tagger features in highly offensive tweets in the humorous subset.