UNIVERSITAT POLITÈCNICA DE VALÈNCIA

etsinf

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

School of Informatics

When humour Hurts: A Computational Linguistic Approach

End of Degree Project

Bachelor's Degree in Data Science

AUTHOR: Merlo , Lucia Ines

Tutor: Rosso, Paolo

Experimental director: CHULVI FERRIOLS, MARIA ALBERTA

ACADEMIC YEAR: 2021/2022

Summary

Traditionally, humour is considered as a funny way of communication. Although it can achieve laughter in the receptor by making use of non hurtful language, it is not always like that. Frequently, humour is applied to address controversial topics, sometimes being hurtful to a person or people who belong to certain groups. Hence, nowadays some researches focus on how hate speech is disguised into humour. The main objective of this study is to use a computational linguistic approach to detect in humorous texts, which are the characteristics that distinguish high and low levels of offence in jokes. Variables detected as relevant in the previous characterisation are applied into a classification model. With this second step, this work analyses how well a Machine Learning model performs, and by applying an ablation test, variables that stand out within the classification task are identified.

Keywords— Humour, offensive language, computational lingüistics

Resumen

Tradicionalmente, el humor se considera una forma de comunicación divertida. Aunque puede provocar la risa del receptor haciendo uso de un lenguaje no hiriente, no siempre es así. Con frecuencia, el humor se utiliza para abordar temas controvertidos, resultando en ocasiones ofensivo para una persona o grupo de personas. Por ello, en la actualidad, algunas investigaciones se centren en cómo el discurso del odio se disfraza de humor. El objetivo principal de esta investigación es utilizar la lingüística computacional para detectar en los textos humorísticos cuáles son las características que distinguen niveles altos y bajos de ofensa en los chistes. Las características detectadas como relevantes se utilizan un clasificador. En esta segunda parte del trabajo se analiza el rendimiento de un modelo de Machine Learning y, mediante la aplicación de un test de ablación, se identifican las características más relevantes en la tarea de clasificación.

Keywords— Humor, lenguaje ofensivo, lingüística computacional

Acknowledgements

To my family and friends which are my principal support.
Also to my supervising professors Paolo Rosso and Berta Chulvi, who trusted in me all the way through this project, and to Reynier Ortega Bueno, who helped me with the development of experiments.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Humour and Offense

The first result yielded when searching on Google the definition of humour is: *"The quality of being amusing or comic, especially as expressed in literature or speech"*. Humour is everywhere. Over the years it has been evolving, depending on each person state of mind, or in a more general way, regarding social contexts within each period. Moreover, being humorous is considered as a quality leaders should have [11]. While it can be used as a method for encourage people, it also might produce a change in a person belief on a certain topic. Moreover, a theory in psychology maintains the fact that humans only find something humorous if the target is interesting for them, meanwhile also depends on the cultural context [15]. However, it always had the same goal: being comic and make people laugh. From daily jargon, passing by TV, streaming platforms and social media, the way of make the ordinary fun has been changing. Comic videos all over streaming platforms, memes going viral in any type of social media as shown in Figure 1.1, are some frequent ways of making humour in our day by day, besides the traditional joke-telling. Another way humour has changed, relies on the things and situations considered funny. Something amusing twenty years ago, nowadays might be considered boring, aggressive or even hateful.

At this point, it is necessary to introduce the concept of *adversarial humour* [38]. Studied in the psychological field of humour, enables to explain up to a certain point social interactions regarding jokes and the underlying goals of them. Opposite agents (adversarial individuals) have interests that can induce, in terms of communication, verbal conflicts due to specific goals pursued by each part. Rivers of ink can address adversarial humour. This research considers this humour branch as nearly related to double-grounded insults, metaphors, role-reversal and competitiveness. Therefore, it can be described as *turning tables* on another person in a conversation context.

As a consequence, offensiveness with the goal of hurt someone can come out, even in a humorous context. Despite this, a joke can have abusive language but not being hurtful. In fact, a joke is assessed as hurtful when it is intentionally directed to a target. Furthermore, hate speech directly attacks or promotes hate to a group or an individual, only considering their identity (ethnics, sexual orientation or religion). Nonetheless a joke can be hurtful without being explicitly abusive [42], in other words, having implicit abusive language being the reason why sometimes offensive jokes can not be properly detected.

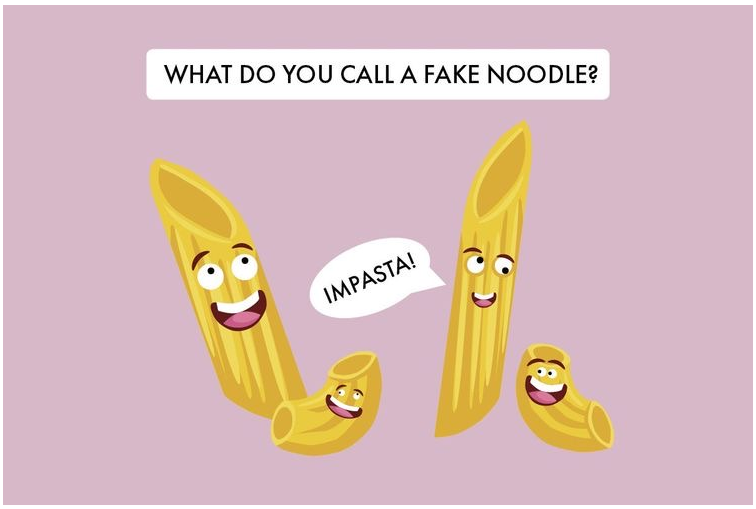


Figure 1.1: Joke extracted from Twitter [1]

1.2 Motivation

The branch of humour recognition within Natural Language Processing (NLP) is not as broad extended as it should. In fact, machines tend to perform better on humour generation, rather than in this task [21]. Although it is considered a key ability in functional communication and as a tool for improving interpersonal relations, what makes people laugh is not deeply studied. However, it is known that people tend to be attracted to what they know and personal experiences [5], something applicable to this context. Draw from these premises, it is relevant to dive into humour from a computational perspective in order to detect situations that could be unnoticed for human eyes. From recognising subtle language patterns in big data, to distinguishing language markers that characterise certain forms of interaction, possibilities are uncountable.

Moreover, the motivation of this project does not rely uniquely on the scientific field. Personally,

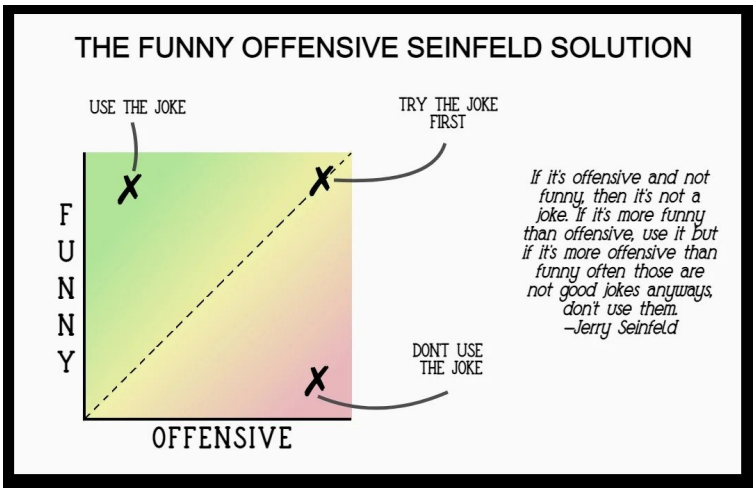


Figure 1.2: Offensive joke solution [12]

I consider NLP a field which enables to combine several tools in order to extract a deeper understanding from data. From a technical point of view, programming is one of the cornerstones of the named field. Data exploration and annotation, building linguistic features, experimentation and model development are some of the tasks which require computer science skills. Furthermore, new methodologies, diving into state of the art and investigate new ideas are also requirements of this research. However, Data Science is not static discipline, as neither NLP. Human language is evolving each day, being fundamental the study of language from its structure as well as from a psychological approach. Human capability of communicating something without explicitly referring to it seems fascinating to me. For this reason, the topic of this project relates to offensive language even when it comes disguised into a funny/comical context.

1.3 Objectives and Research Questions

As previously referenced, humour has always been considered as a fun form of communication. Nevertheless, it is well known the existence of humour with an underlying hurtful attitude. Nowadays, social media platforms are widely extended all over the world, a situation that often gets the most aggressive and antipathy of people. Twitter is one of the platforms mostly used all over the world, for expressing points of view with impunity and anonymity, and it compose a great space for hate speech and aggression. This type of expressions/words/phrases are often camouflaged into jokes, trying to hide underlying negative attitudes. This yields the interest on analysing if there are patterns present in hurtful humour, that could help to the identification of these type of communication. As a consequence, four research questions arises:

- RQ1. Which are the features that distinguish humorous texts from the non humorous ones?
- RQ2. Which are the features that distinguish non offensive humour from the offensive one?
- RQ3. How do classifiers perform distinguishing non offensive humour from the offensive one?
- RQ4. Which are the characteristics that enable classifiers to make this distinction?

1.4 Thesis Structure

This document is organised as it follows: in Chapter two 2 a brief insight on theorethical framework until nowadays, a description of ethical and legal analysis and which are the aspects of this field improved by this study are done. Description of the dataset, resources applied and followed methodology can be read in Chapter 3. In Chapter 4, analysis of relevant features is carried out for both, humour and offensive humour detection. In Chapter 5, a classification task is computed where results and discussion. Lastly, in Chapter 6, we address the conclusion of the findings, go into the contribution of this study, the relation between the final project and Data Science degree, as well as the discussion of future work. For organisation aims, further results are available in appendices A, B and C. In appendix D a reflection, linking this project with the sustainable development objectives is addressed.

Chapter 2

State of the Art

2.1 Introduction

In the last decades, it has been possible to observe the exponential grow of social media worldwide. If in December of 1995 there were 16 million of users, nowadays more than 5,000 million of people are navigating through the Internet [14]. Through this tool, people in any geographical point, are able to talk about a diverse amount of topics, from politics, health, society, business, science to entertainment.

This intensive use of social media is closely linked to humorous manifestations: people are capable of making fun of anything in a healthy or offensive way. The use of humour to offend a particular person or a group of people, is something extremely easy to do with the anonymity that social media supply. Hence, some studies as [42], relate that any kind of subject can be approached either from a funny way and, at the same time, in a hurtful way. Offensive language that offends minorities can be disguised inside humour and stay anonymized inside the web.

In this chapter it is shown how the interest over the detection of humour has been increasing in the last years in computational science research. Moreover, it is commented how the majority of investigations focus on detecting humour as a classification task in a computational approach. However, an increasing number of researches have lately set out to take a step further regarding humour recognition. With this aim, most of them try to find which mechanisms (semantic incongruity, irony, idioms) are applied to achieve humour. Moreover, another objective tackled consists in detecting the humour target, as well as the purpose of it (hurt, offend).

As it is referenced in this chapter, due to the fact that targets of prejudice (likewise ethnic or gender groups) are frequently referred in jokes, a good part of research in this study field is related to the recognition of hate speech within social media.

2.2 Theoretical Framework

One of the first analysis for humour detection considering linguistic features was presented [21]. Authors considered a different approach from the traditional classification task, focusing uniquely

on model accuracy. They carried out a study for distinguishing humorous and non humorous texts, using a computational approach for humour detection. Furthermore, humorous examples consisted in one-liners while non-humorous texts were extracted from three resources: Reuters news titles, proverbs and texts from British National Corpus (BNC). In English context, one-liners is a 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, systems were trained with stylistic features (alliterations, antonyms and slangs), content features and a combination of both. The results obtained showed that stylistic marker, help to distinguish a large number of one-liners between Reuters titles, while proverbs text style seemed to be similar of one-liners. A similar case for one-liners and news titles was detected for content markers, meanwhile jokes were identified as similar regarding BNC. Moreover, authors suggest that content is similar for jokes and normal texts. Nevertheless, content features help to differentiate jokes and proverbs, although their stylistic similarity. This study reveals that it was tough for systems to distinguish jokes from normal texts. However, authors remark how humorous data mostly refer to human-scenarios likewise words referring to an individual (man, woman, I, you, person), besides negative forms of words (isn't, doesn't, bad). Lastly, from the observation of the learning curves, researchers suggest that sophisticated linguistic features could improve texts classification, instead of data augmentation.

Analysis of humour done from a linguistic and human-centric content approach [30] was studied for online comments. This study had the aim of improving humour detection within the Web, in order to increase entertainment value of visitors of a certain a Web page and distinguish implicit humour and the absence of it. A selection of human-annotated funny comments and one-liners from a news website, and experiments through classification models were carried out. These ones were trained with linguistic features belonging to: sexual terms, terms with negative polarity, semantic ambiguous terms, affective terms, or slangs plus emojis. By computing a multi-label classification either in funny, informative, insightful or negative category, authors examined which of the features applied and contribute the most to humour detection. They found that funny and informative categories were similar, while insightful and negative ones seemed to be distinct. Moreover, regarding the features, slangs terms and emojis helped to improve jokes detection, unlike affective features for the one-liners case. Furthermore, this article mention the good results on one-liners, oppositely on comments. The main explanation relies on the fact that negative comments present the same structure as funny ones. Moreover, authors observed that humour in one-liners is caused mostly by the use of irony, sarcasm and ambiguity. Whereas humour within comments can be reached out as responses, introducing discrepancy. Being the last point a possible cause in the lack of performance of features.

The construction of systems for humour recognition applying a simplistic approach can provide intriguing results. Taking this statement as starting point, authors of [34] sought to determine whether a text is a joke or not without considering the meaning. They addressed this issue putting the focus on identifying which linguistic features could be useful for joke detection. The features considered were construct in a shallow way, with the purpose of evaluating if classifiers could perform correctly solely with readily available and simple variables, instead of more sophisticated ones. These variables were extracted from a corpus of 6,100 one-liner jokes and phrases from the British National Corpus for non-humorous examples. As a result, all texts are in the English language.

The features considered are: text similarity (word overlap between the training instance and 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). Subsequently training the classifiers, this study yielded as a result that common and rare terms seemed to be useful for humour distinction. Moreover, word overlap, idiomatic expressions and ambiguity are groups of variables that compose this contribution. On the other hand, stylistic features did not seemed to provide a substantial contribution.

Authors ground their conclusions on the fact that word overlap and detection of a particular set of words helped the systems in humour identification. Besides, this article addresses how humorous texts differ from others depending on the approached topic. This conclusion goes in line with the fact that features extracted from content markers are highly relevant. Lastly, the development of more sophisticated features from those applied in the study and considered as non relevant, was suggested as future work. Likewise, variable related to word repetition, as it may be relevant for only few texts and drowned out by remaining jokes, at least in this dataset [34] (e.g. *"Kids in the back seat cause accidents; accidents in the back seat cause kids."*).

Lastly, it is convenient to remark that humour can come in diverse forms, and memes are one of them. Memes are images or pictures with text on them, with the goal of reflecting everyday situations in a humorous tone. Nevertheless, these sometimes contain subjective jokes related to several topics (politics, religion, ethnics, sexuality) and moreover, they may be hurtful. These content can be quickly spread and reach millions of people, by the way of social media. As a consequence, people's opinion can be influenced by them. Hence, authors of [10] suggest the necessity of developing systems which detect if a meme has offensive content in it, before going viral. Based on a dataset of 6,992 memes, the methodology followed consisted in converting images to text and afterwards applying text preprocessing. As Word embeddings they employed *GloVe* and *FastText*. By means of Neural Network models, researchers tackled two classification tasks. First one was to detect whether the text within a meme was offensive or not. In case of being offensive, as a further step, a second classification was done in three categories: slightly offensive, very offensive and hateful offensive. Data augmentation was done by authors in order to avoid a possible overfitting problem. The model trained with *FastText* was identified as the most efficient in terms of time execution, as well as the one which got better results. The study was developed with a predictive approach, without adding a descriptive point of view. Authors recognised the possibility of implementing extra feature engineering previous to examine which are the characteristics that compose a offensive meme, as it could be extrapolated from the topic of this final project.

Although referenced studies cover in some way which features can have greater or less impact on humour, it still exists a gap regarding studies that describe the relation between humour and offense and the impact of linguistic markers on their detection. *SemEval 2021 Task 7, HaHackaton, Detecting and Rating Humor and Offense* [20] is the first shared task that copes with the detection of both, humour and offense. The submissions received by the organisers are from a Machine Learning point of view, showing a predictive insight only.

There were four subtasks: two classification tasks (humour and humour controversy detection) and two regression tasks (humour rating and offense rating detection). For their evaluation organisers developed two benchmark systems. Using *sklearn*, for classification tasks, Naïve Bayes model was generated with Bag of Words (BoW) features. For the regression tasks, Support Vector Machines (SVM) with term-frequency inverse document frequency (tf-idf) features was employed. However, the organisers also employed for the classification and regression tasks, BERT-based system. As this model outperformed in all tasks, with a F_1 measure of 0.9 for humour detection, and 0.4 for offense rating, it was selected as baseline. A total of 63 teams submitted systems for distinct tasks. The metric selected for evaluating each submission was F_1-score, and in case of a tie, accuracy of models was considered. Teams with the best results applied model ensembling by majority voting in classification tasks and average in regression tasks. The most popular systems were pre-trained models likewise BERT, ALBERT and RoBERTa, among others.

These methods resulted efficient for most tasks, however, humour controversy remained particularly difficult to detect. After carrying out a deeper analysis of the data and benchmark systems, authors of [20] reference the existence of a negative correlation between the detection of offense and humour, plus a particular type of humorous texts that several powerful models struggle to capture.

2.3 Proposal

Taking into account the state of the art, it is interesting to address the relation between humour and offense as a classification task and from descriptive perspective, by studying which type of variables help to distinguish humour from non humour. As a step further, with the implementation of an ablation test, it is proposed to analyse which linguistic markers contribute the most to discriminate between non offensive humour and offensive one. This research can make meaningful contributions from computational linguistics to fields related to social and political science which try to detect when hurtful speeches are being promoted using humour, specifically on social media. Identifying this offensive humour is a first step in order to provide a good type of response and avoid a quick proliferation of hate through the Web.

2.4 Problem Analysis and Legal and Ethical Analysis

At the present time, all information inside the Internet own its authorship to someone. Specifically on social media, where all users must be adequately registered in order to make use of these platforms. Lately, companies, governments and organisations have detected data as a key point for their own development and improvement. However, these data become something deeper than a chucklesome username, it is generated by regular people with their own issues and worries. Data protection has become more relevant than ever, where privacy is one of the cornerstone of this field.

This project makes use of data extracted from social media, which is necessary for assessing this project. Concretely, the project makes use of data provided by [20], which takes part of *SemEval-2021 Task 7*. The named dataset contains texts extracted from Twitter. However, the organisers of this competitions took particular care of the tweets collection. The namely texts belong to ac-

counts destined as public forum (e.g. @BadJokeCat,@ZaraRahim). As a result, sources accounts are publicly available [20]. Despite this, when constructing and processing the dataset, the organisers removed usernames and any word/hashtag/link from tweets in order to make texts completely anonymous.

In terms of ethics, this study is important because the data used could cause a detrimental on groups of people. In fact, a several amount of analysed texts make reference to sensitive topics, majority of them refer to a group of people (Mexican people, Asians), an individual that has certain characteristics (black person, Jewish man/woman, fat woman), or someone popular (Kim Jong-Un, Kim Petras, Kim Seltzer) who is denigrated. Therefore it could affect in a negative way either social groups (e.g. minorities) or someone in particular.

Chapter 3

HAHA Task, Data Resources and Methodology

3.1 HAHA Task Dataset

For this research it is used the *HAHA* Task Dataset from *HaHackaton* organised at IberLEF 2021 [20]. IberLEF is the workshop on Iberian Languages Evaluation Forum of the SEPLN conference. To construct this dataset the organisers extracted data from the Internet, in English. Specifically, 80% of texts are originated in Twitter and unsettled 20% is obtained from the Kaggle *Short Jokes* dataset [24]. The purpose of getting a portion of data from a previous constructed dataset, is to ensure humour quota (texts intended to be humorous), traditional humour quota (texts conventionally recognised as humorous) and offense quota (texts intended to be offensive).

As referenced in the *HaHackaton overview* paper [20], if a keyword related to offensive terms is contained in the text and is the target of the joke, then it is possibly an offensive text. In contrast, if the text has the keyword in it but it is not the target of the joke, is considered as non offensive, likewise in Table 3.2. Some examples of these terms are shown in Table 3.1.

Table 3.1: Examples of targets and keywords mentioned in the HAHA overview paper [20]

Target	**Keyword**
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 3.2: Examples of jokes with keywords mentioned in the HAHA overview paper [20]

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

Regarding Twitter data, humorous texts have been selected from some accounts (e.g. @JokesMemesFacts, @Dadsaysjokes, @BadJokeCat). For non humorous content, selected accounts post about news, celebrities, organisations and quotes (e.g. @CNN, @Oprah, @BlkMentalHealth).

The task organisers crawled 2,000 tweets from each account with Twitter API. Texts with links, US politics content, COVID-19 and retweets were removed, while hashtags within tweets were substituted with the constituent word. As a final result, a total of 8,000 tweets compose the dataset. In order to avoid biases regarding linguistics, texts and human annotators origin were entirely from the United States.
Text annotation was done by US citizenship participants, each one belonging to one age group (18-25, 26-40, 41-55, 56-70). Each text was annotated by 5 members of each group. Annotations marked if the text intended to be humorous, if it is generally offensive and if it is personally offensive. If the reader replied affirmatively to any of these questions, that person had to evaluate the content of the text to determine if it was intended to be humorous, and rate the text with an humour and offense score from 1 to 5. Furthermore, in this task, it was considered uniquely general offense annotations. Kaggle texts were used as data quality control over tweets annotations, in order to evaluate if voluntaries followed the annotation instructions properly. Texts were labelled as humorous if more than half of annotations were marked as that. In case of a tie, texts were assigned humorous label. Humour and offense rating were calculated as the average of ratings assigned to each text. For texts classified as non offensive (i.e. "no"), the given rating was 0. The same procedure was taken for humour rating, having the same range as offense rating.

3.2 Linguistic Resources

This section describes the linguistic resources used to carry out the texts analysis.

Stanza

The Stanford Natural Language Processing Toolkit [19], also known as Stanza, is a tool for Natural Language Processing analysis, a Python library, that provides a detailed breakdown from human written texts. That is, it splits a text into lists of sentences and words. On the basis of this lists, it supplies a deeper insight into the text analysed. By the development of Neural Networks models, this library enables several features such as tokenization, lemmatization, part of speech and morphological tagging, as well as dependency parsing, entity recognition, language identification.
Furthermore, Stanza through an interface which applies a Convolutional Neural Network (CNN) classifier, it enables texts sentiment analysis.

SenticNet

SenticNet [6] is an adaptation of SenticNet4 presented as a Python interface applied in Sentiment and Semantic Analysis. This tool extracts affective information from human language, in order to analyse it quantitatively in a computer friendly way. By a multidisciplinary approach which includes mathematics, statistics, psychology and linguistics, SenticNet addresses text analysis from two points of view. In order to extract meaning, semantic networks and dependency parsing are generated from a string. Meanwhile, deep neural network models are computed for extracting patterns underlying data. With 400,000 natural language concepts as a foreground, this tool applies deep learning to extract from words and phrases, primitives and subsequently into superprimitives, as shown in Figure 3.1.

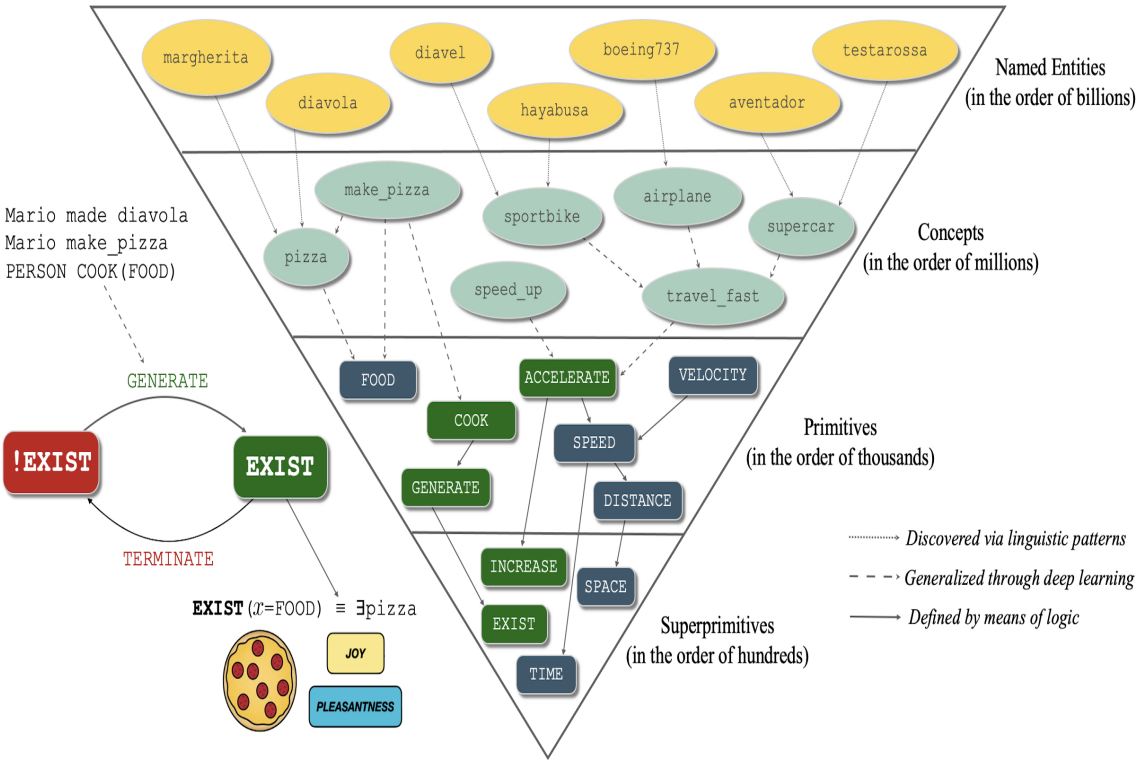
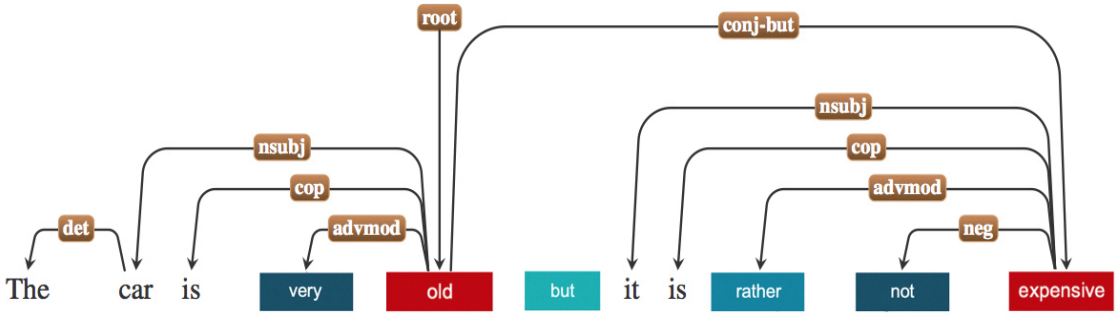
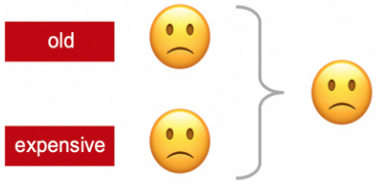


Figure 3.1: Sentic baseline [32]

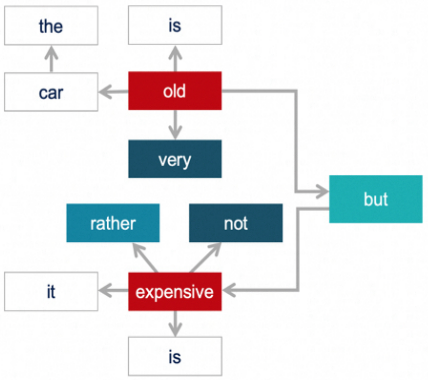
Hereby polarity within texts can be brought into light via *sentic patterns* [29]. With this baseline, a sentence can be studied by applying these patterns to its dependency tree. The extracted information, is the result of considering core concept terms. That is, words that change the meaning of the phrase and elements without polarity associated. Lastly, the resultant tree can be analysed as a electronic circuit as shown in Figure 3.2.



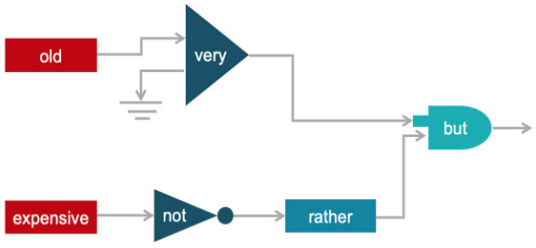
(a) Dependency tree of a sentence.



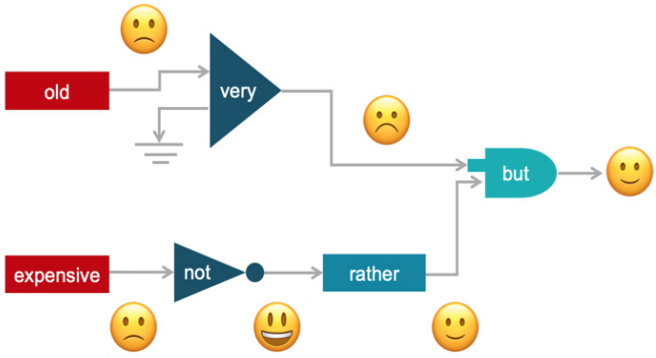
(b) The old way: averaging over a bag of sentiment words. The overall polarity of a sentence is given by the algebraic sum of the polarity values associated with each affect word divided by the total number of words.



(c) The dependency tree of a sentence resembles an electronic circuit: words shown in blue can be thought as a sort of "boolean operations" acting on other words.



(d) The electronic circuit metaphor: sentiment words are "sources" while other words are "elements", e.g., *very* is an amplifier, *not* is a logical complement, *rather* is a resistor, *but* is an OR-like element that gives preference to one of its inputs.



(e) The final sentiment data flow of the "signal" in the "circuit".

Figure 3.2: Sentic dependency syntactic tree [32]

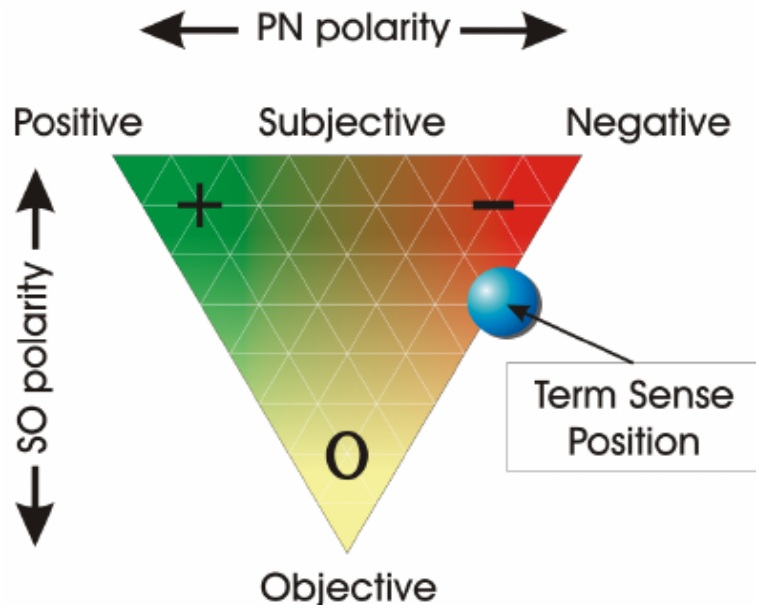


Figure 3.3: SentiWordNet graphical representation [9]

TextBlob

TextBlob is a Python tool developed to process text data, supported by *Natural Language ToolKit* and pattern, both developed for Natural Language Processing tasks. Similarly to the Stanza tagger, this tool enables tokenization, lemmatization, part of speech and morphological tagging, n-grams extraction, words and phrase frequencies, classification, and more. TextBlob is a lexicon-based tool. Each word has a weight associated within the lexicon. As described in [37], it is called "Rule-based sentiment analyser", given that each word's weight hints polarity computation.

SentiWordNet

SentiWordNet [9] is a lexicon resource for opinion mining derived from the WordNet database, where both make use of synsets. Synsets are nouns, adjectives, adverbs and verbs grouped as a set of synonyms. Each synset has a concept associated and each term composing the synset is a lexical variant of that concept. For instance, the concept "media" can have a set of terms associated to it, such as radio, television, magazines, and internet. In the case of SentiWordNet, SO-polarity (Subjectivity and Objectivity polarity) and PN-polarity (Positive and Negative polarity) are determined for each synset as shown in Figure 3.3.

Specifically, this tool consists in a lexical resource with a total of 17,530 synsets. These are represented as vectors, obtained by applying cosine normalised with *tf*idf* (term-frequency inverse document frequency), preceded by removing stop words. Each synset has three scores associated. These scores reflect the level of positiveness, negativeness and objectiveness of each term within the synset. Each score ranges from 0 to 1, and for each synset their sum must be equal to 1.

VADER

Valence Aware Dictionary and sEntiment Reasoner [13] is a lexicon and rule-based sentiment analysis tool designed for social media sentiment analysis. Python code for rule-based sentiment analysis engine applies grammatical and syntactical rules described in [13]. While being sensitive on the word-order within each phrase analysed, it also considers the presence of intensifier terms when computing the intensity of the sentiment score. On the subject of scoring, two type of score are provided. The first type is associated with positive, neutral and negative scores, which correspond to ratios. These ratios are calculated as the proportion of each term classified as positive, negative and neutral, giving a detailed picture of the way sentiment is distributed while providing context. On the other side, second type of score remains in compound scoring. Presented as the sum of valence score of those words present in the VADER lexicon and normalised in the range [-1;1] where -1 is most negative and +1 most positive sentiments, provides an overall picture regarding the text study.

ANEW

Affective Norms for English Words [39] is a lexicon with 14,000 terms in their lemma format manually rated by volunteers on valence (pleasure), arousal (excitement), and dominance (level of control) measures.

AFINN

AFINN is a lexicon-based Python analyser and a lexicon by itself. As a variant of the ANEW resource, it was developed from tweets for sentiment analysis. The word list presents 2,477 words, each one labelled with a valence score between [-5;5], being -5 very negative and +5 very positive sentiment. Annotations were done manually by the author. The main content is a list of obscene words as well as few positive terms. The paper [26] presents a detailed background of how it has been constructed. Moreover, in the paper the author refers to the existence of a bias related to negative terms.

Lexicon of Abusive Words

This lexicon composed by Hate Speech related terms developed by [41], has two lexicons types. The base lexicon was built with 1,650 words manually labelled via crowdsourcing. Each word was tagged with “TRUE” for abusive terms and “FALSE” for non abusive words. Words labelled with “TRUE”, were considered as abusive if 4 from 5 annotators coincided. On the other hand, the expanded lexicon consists of 8,947 words with a score assigned to them, by applying a classifier to each term. The binary lexicon has been considered as a baseline for constructing this resource. Positive scores are related to abusive words, while negative scores to non abusive terms.

Hurtful Words

HurtLex [4] is a lexicon with offensive words classified in one of the 17 categories shown in Table 3.3 present in the resource. Moreover, each term and category label are enclosed either in conservative (words with explicit hate) or inclusive (words with potential hate connotations).

Table 3.3: Hurtlex categories

Category	Description	Length
PS	Negative and Stereotypes and Ethnic Slurs	371
RCI	Location and Demonyms	24
PA	Profession and Occupation	192
DDP	Phsyical Disabilities and Diversity	491
DDF	Cognitive Disabilities and Diversity	63
DMC	Moral and Behavioural Defects	715
IS	Words Related to Social and Economic Disadvantage	124
OR	Words Related to Plants	177
AN	Words Related to Animals	996
ASM	Words Related to Male Genitalia	426
ASF	Words Related to Female Genitalia	144
PR	Words Related to Prostitution	276
OM	Words Related to Homosexuality	361
QAS	Words with Potential Negative Connotations	518
CDS	Derogatory Words	2204
RE	Felonies and Words Related to Crime and Immoral Behaviour	619
SVP	Words related to the Seven Deadly Sins of the Christian Tradition	527

This resource has been constructed through another resources. Firstly, the lexicon called "words to hurt" *Le parole per ferire*, presents more than 1,000 terms in Italian which belong to one of three distinct categories: derogatory words, words related to stereotypes and words not explicitly hurtful but can be used with that purpose. Secondly, an augmentation of words with their part of speech tags took place by applying MultiWordNet [33] (an extension of the WordNet lexicon) and adding identifiers of synsets of each term lemma, showing if it is detected as a verb, noun, adjective or pronoun with BabelNet [25]. Subsequently, a manual annotation was done in order to discard any type of non offensive term.

EmoSenticNet

EmoSenticNet [3] is an extensive lexical resource, developed to assign emotion labels to words and phrases while containing polarity scores of each term. The main goal remained in extracting the information related to a settled emotion, meanwhile knowing the orientation of sentiment (polarity) related to that term. That is, a combination between SenticNet [6], WordNet-Affect [35] resources and the ISEAR dataset [31] has been used for the construction of this lexicon. By training different classification models and varying their hyperparameters, the final result consists in an assignation of each word into one of the six emotion labels, as shown in Table 3.4.

Table 3.4: EmoSenticNet categories

Category	Description	Length
Anger	Concepts related to anger	828
Disgust	Concepts related to disgust	1158
Fear	Concepts related to fear	1198
Joy	Concepts related to joy	9388
Sad	Concepts related to sadness	1535
Surprise	Concepts related to surprise	904

SentiSense

Similarly to EmoSenticNet, SentiSense [8] is a concept-based affective lexicon based on WordNet. Containing words and synsets labelled with a determined emotion, it has been constructed in two steps. Firstly, two annotators labelled each word and concept. Subsequently, the result obtained by the annotators was expanded by applying semantics relations between concepts present within WordNet.

Table 3.5: SentiSense categories

Category	Description	Length
Anger	Concepts related to anger	54
Anticipation	Concepts related to anticipation	151
Disgust	Concepts related to disgust	547
Fear	Concepts related to fear	159
Joy	Concepts related to joy	132
Like	Concepts related to like	345
Love	Concepts related to love	55
Sad	Concepts related to sadness	134
Surprise	Concepts related to surprise	29

LIWC

Linguistic Inquiry and Word Count is a dictionary-type resource [36] widely used in the interdisciplinary field that links psychology with computational linguistics. It is focused on the social and psychological meaning of words in order to capture people's cognitive styles and emotional states. As a result, this resource contains terms associated to a wide set of categories, often ordered in a hierarchical format. It includes more than 100 dictionaries, where each one contains words, emoticons, word stems and verbal constructions.

Table 3.6 shows the categories contained in LIWC 2015. Some of them are linguistic (function words, part-of-speech terms), psychological (affective processes, emotions, social processes), cognitive, perceptual, biological processes and more.

Table 3.6: LIWC 2015 [36]

Category	Feature	Example	Length	Category	Feature	Example	Length
Linguistic processes				Psychological processes			
	Function words	Of	464		Perceptual processes	Observe	273
	Pronouns	Itself	116		See	View	72
	Personal pronouns	Them	70		Hear	Listen	51
	1st Person Singular	I, me	12		Feel	Touch	75
	1st Person Plural	We, us	12		Biological processes	Eat	567
	2nd Person	You, your	20		Body	Hands	180
	3rd Person Singular	She, her	17		Health	Flu	236
	3rd Person Plural	They	10		Sexual	Horny	96
	Indefinite Pronouns	It	46		Ingestion	Eat	111
	Articles	A, an, the	3		Relativity	Area	683
	Verbs	Went	383		Motion	Car	168
	Auxiliary Verbs	Am, will	144		Space	Thin	220
	Past Tense	Went	145		Time	End	239
	Present Tense	Is, do	169	Personal concerns			
	Future Tense	Will	48		Work	Job	327
	Adverbs	Very	69		Achievement	Earn	186
	Prepositions	To	60		Leisure	Cook	229
	Conjunctions	And, but	28		Home	Apartment	93
	Negation	No, never	57		Money	Cash	173
	Quantifiers	Few, many	89		Religion	Church	159
	Numbers	Second	34		Death	Kill	62
	Swear Words	Damn	53	Spoken categories			
Psychological processes					Assent	Agree	30
	Social processes	Talk	455		Nonfluencies	Er	8
	Family	Son	64		Fillers	yaknow	9
	Friends	Buddy	37				
	Humans	Adult	61				
	Affective processes	Happy	915				
	Positive emotion	Nice	406				
	Negative emotion	Hurt	499				
	Anxiety	Nervous	91				
	Anger	Annoy	184				
	Sadness	Cry	101				
	Cognitive processes	Cause	730				
	Insight	Consider	195				
	Causation	Because	108				
	Discrepancy	Should	76				
	Tentative	Maybe	155				
	Certainty	Always	83				
	Inhibition	Stop	111				
	Inclusive	Include	18				
	Exclusive	But	17				

3.3 Statistical Resources

Statistical methods applied in this research are explained in this section. Parametric statistics are used when selected data fulfil certain assumptions on their distribution. For cases where these assumptions are not reached, non-parametrical statistics [23] are a satisfactory alternative.

Non-parametrical statistics are applied when it is not possible to assume that variables follow a normal distribution, as a consequence, are acknowledged as "distribution-free" methods. In cases where the sample size is not as large as based in the Central Limit Theorem [16] and data has an unknown distribution, the application of these types of methods are also appropriate. Statistical analysis enables to study the certainty of postulations from a set of assumptions and the sustainability of a conclusion. As long as the proper method is applied, incorrect or skewed results will be avoided. The main objective of statistical tests is to determine if a theory about a process or event has to be either rejected or accepted, where it is possible to apply parametric or non-parametric tests. Most of the data used in this research have an unknown distribution. Non parametric analysis used in this research is explained in the following paragraphs.

Spearman Correlation

Spearman correlation is a statistical measure which reflects the monotonic relationship between paired data. It takes under consideration variables which contain ranks, ratios, interval or ordinal values. This measure is known as *rho* or r and it ranges between [-1,+1]. If the absolute value of r is near to 1, it mean that the association between variables is strong. When the value of r is close to 0, the relation between variables is low or null.

Mann-Whitney U Test

The Mann-Whitney U test is a non parametric method, which compares if two independent groups present significant differences between them on a dependent variable. This test must follow certain assumptions for drawing proper conclusions. Firstly, the dependent variable must be measured in either ordinal or continuous scale. For instance, ranking categories (e.g. strongly disagree, disagree, agree, strongly agree) or quantitative (e.g. number of hours, salary, weight or height). Secondly, the independent variable must be composed by two categorical groups (e.g. men or women, smoker or no smoker). Furthermore, the observations analysed must be non paired. The goal is to analyse independent samples within and between groups in order to avoid biases. Lastly, the Mann-Whitney test can be applied in non normal distributed data.

Wilcoxon Signed-Ranked Test

The Wilcoxon Signed-Ranked test is a non parametric method which compares if two groups of paired data present significant differences between them on a dependent variable. Similarly to the Mann-Whitney U test, dependent variable must be either ordinal or continuous, while the independent variable must be composed by two categorical groups. Furthermore, it does not assume normality in the data.

3.4 Data Preprocessing

For data preprocessing, the main focus relied on *cleaning* each tweet by applying the Stanza tool [19]. Firstly, a pipeline has been built. This must indicate the specific NLP tasks (tokenization, part-of-speech tagging, lemmatization, and more) to meet. The pipeline receives as input raw text, executes the specified NLP tasks and retrieves an annotated document. After building the pipeline for the English language, each instance is tokenized and converted to lower case. Subsequently, punctuation symbols and non alpha characters within each tweet are removed. Lastly, for each token its lemma is allocated.

3.4.1 Features for Exploratory Data Analysis

This study makes use of two annotation outcomes, "Is humour" and "Offense rating" variables, as the main focus relies on offensive humour detection. Nevertheless, the "Humour rating" feature also takes part of data exploration task. "Is humour" is a binary feature composed by 4,932 positive cases ("Is humour" = 1) and 3,068 negative cases ("Is humour" = 0). Examples tagged as humorous and non humorous are shown in Table 3.7.

Table 3.7: Tweets examples

	Text
Humour	Roses are dead. Love is fake. Weddings are basically funerals with cake.
No Humour	A dog in Mexico named Frida saved the lives of 12 people who were trapped under rubble after an earthquake in 2017. She has identified a total of 52 bodies throughout her career and is considered a national heroine. She's officially retired.

As commented in Section 3.1, "Offense rating" is generated by groups of human annotators and ranges from 0 (least offensive) to 5 (very offensive) (Figure 3.4). Hence, a key step in this analysis remained in tackling if "Offense rating" variable has an expected behaviour whereas being relevant to this study. Firstly, the distribution of this variable is inspected for humorous and non humorous instances. It is noticeable that offense rating variable is right skewed, exposing that higher score values of this feature are less frequent. For the humorous and non humorous breakdown, both seem leptokurtic centred in 0. Nonetheless, humorous data present a higher variance in comparison to second subset. Taking a step further, it is relevant to notice that offense rating ranges from 0 to 4.85 in humour ("Is humour = 1"), while for no humour ("Is humour = 0") the score is between 0 to 3.65.

Pursuing the study objective, statistics of offense rating are inspected uniquely for humorous texts, where some examples are shown in Table 3.8. Four groups are generated, each one corresponds to a quartile of the offense score variable. As it is observed in Table 3.9, groups are balanced. The number of instances within them ranges between 1,264 and 1,205 texts. In addition, the first three subsets have an average score minor to one, while the fourth group presents an average value greater

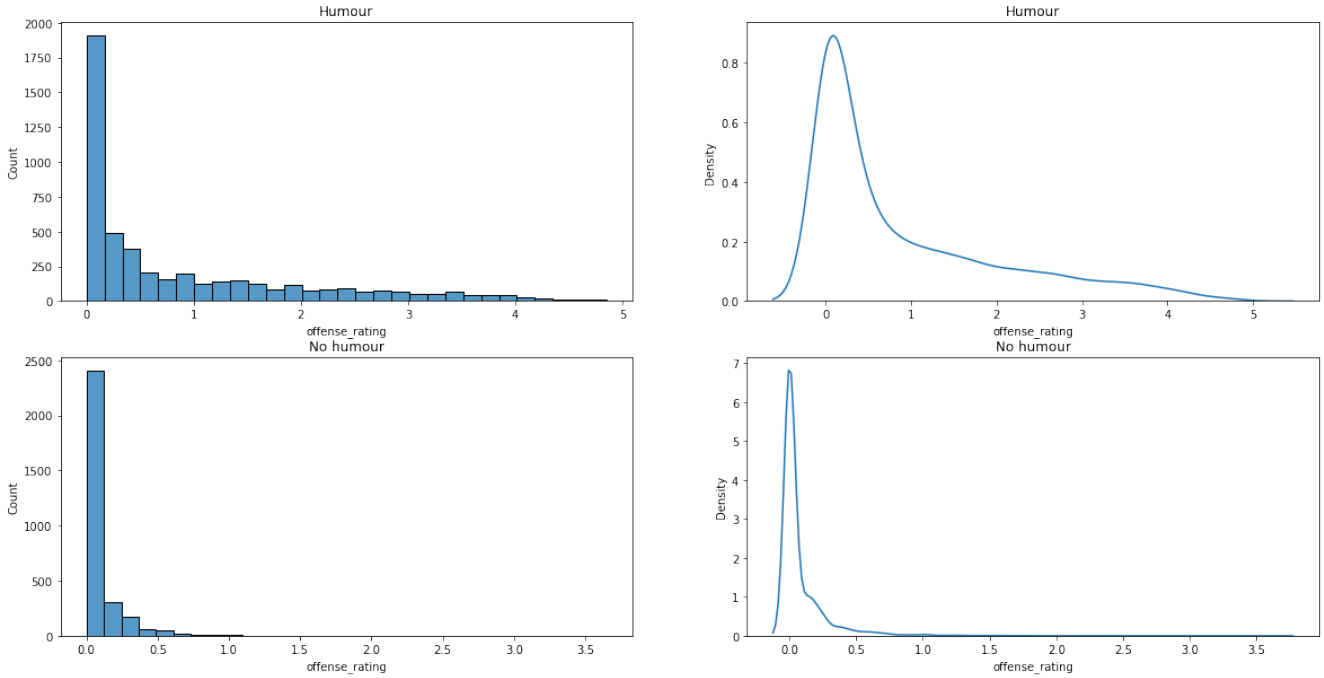


Figure 3.4: Offense rating distribution

than 2.6. As a result, it can be inferred that the first groups are near in terms of offense levels, while texts contained in the fourth set are highly offensive, although the deviation of this group is also quite elevated. Moreover, the Spearman correlation between offense and humour rating scores over humorous data subset is calculated. The objective of it relied on observing if offensive texts were considered highly humorous, although their offensive-related content. With a ρ of -0.27 and a p-value (level of significance) < 0.001, it is verified that the task annotators tend to consider a text with greater amount of humour if the level of offense in it is low or null, as observed in both distributions in Figure 3.5. In other words, as shown in Figure 3.6, the relation between highly offensive content and humour tends to be inversely proportional within HAHA task.

Table 3.8: Humorous tweets

	Text
Non offense	January is the Monday of months
Low offense	I had a vasectomy so I won't have kids But when I got home, they were still there.
Medium offense	What type of wife always knows where her husband is? A widow.
High offense	Black women make the best wives. You can't see their bruises.

Table 3.9: Offense rating groups in humour set

	Non offense (Quartile 1)	**Low offense (Quartile 2)**	**Medium offense (Quartile 3)**	**High offense (Quartile 4)**
Number of samples	1253	1264	1205	1210
Mean	0	0.177	0.846	2.635
Standard deviation	0	0.096	0.322	0.828
Range	[0]	[0.081 , 0.273]	[0.524 , 1.168]	[1.806 , 3.46]

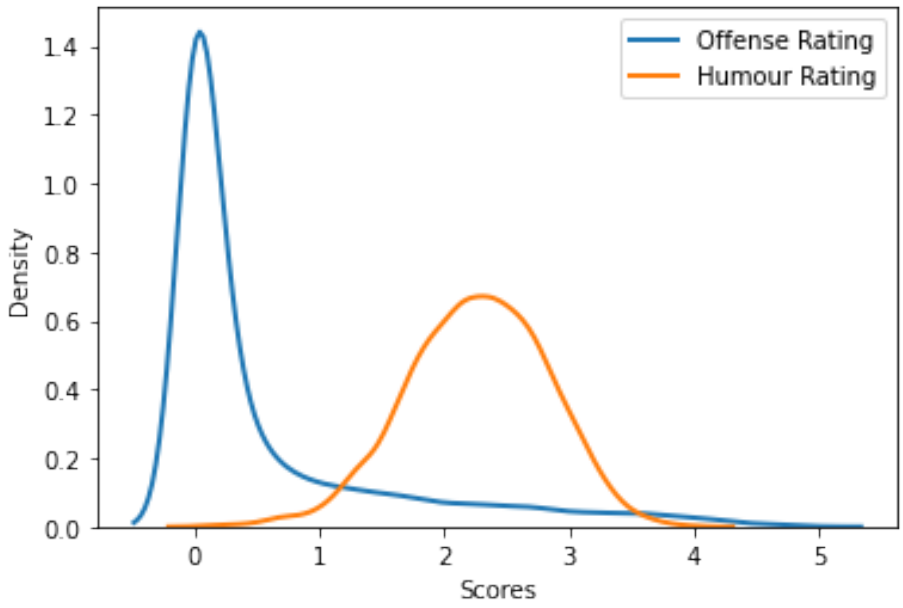


Figure 3.5: Humour and offense rating distribution

Given that part of the study is based on groups composed by non and high offensive instances for humorous set, the following exploratory data analysis is displayed from this point of view. For verifying the quality of annotations regarding offense rating variable, features extracted from taggers are applied to exploratory analysis of this feature. The objective is reached in two steps. The first one consists in computing scoring features calculated from Python taggers, described in Section 3.2.

SenticNet

The feature extracted with SenticNet4 Python API, consists in the polarity score computed from an input text. By creating a class object from Sentic tool called *SenticPhrase*, this can be used for being applied over document. The Sentic feature is a continuous variable, ranging from -1 to +1, being -1 very negative and +1 very positive sentiments.

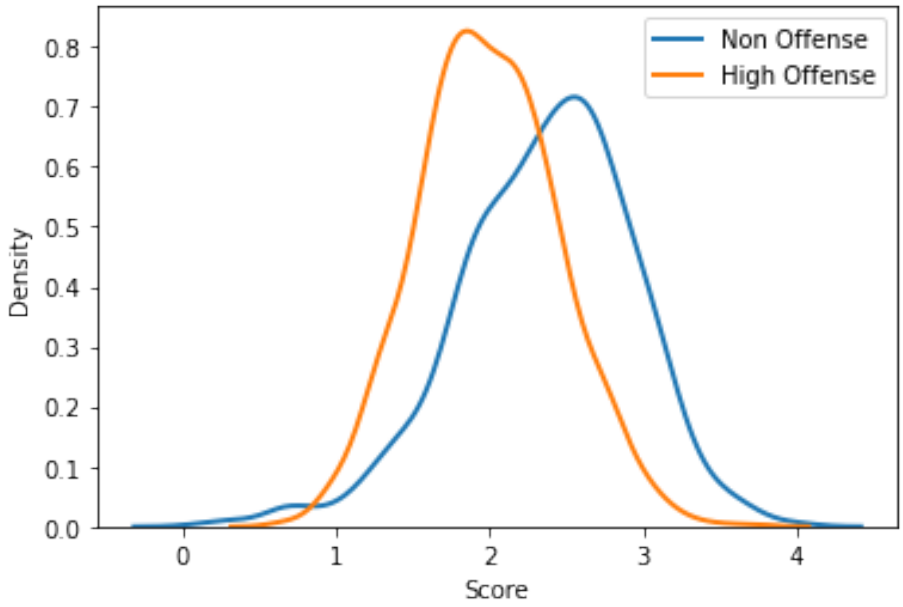


Figure 3.6: Humour rating in the 1st and 4th quartiles

TextBlob

This tool provides polarity and subjectivity of a text. Polarity score range between [-1;1], where -1 is very negative sentiment and +1 very positive sentiment. On the other hand, the subjectivity score ranges between [0;1], quantifying the amount of personal opinion and factual information contained in the sentence. The greater the score, the lower factual information the sentence has.

SentiWordNet

Each tweet is tokenized and part-of-speech tagged. After tagging each term, it is extracted the first synset associated, which corresponds to the most common one. From the selected synset, positive, negative and objective scores are extracted. For each tweet positive and negative scores are considered, where negative and positive scores of each terms synset within a sentence are aggregated. As a result, the final score is obtained by extracting the negative accumulated score from the positive accumulated score, for each sentence. Consequently, for the referenced variable, a MinMax normalisation is computed. Values nearer to 0 are associated to negative sentiments, while those nearer to 1 have a positive sentiment associated.

VADER

VADER tool has been applied for extracting the compound score over original tweets. This score is the sum of the valence score associated to each word, and ranges between [-1;1], where -1 is extremely negative and +1 extremely positive.

ANEW

Three distinct features have been extracted from the ANEW resource. Each term from the lexicon has associated scores of valence, arousal and dominance. Data annotated was split into groups, regarding their age, gender, and educational level. These ratings range between [1;9], where 1 corresponds to least valence/arousal/dominance, 9 the most of them and 5 neutral. For each instance, a score average is computed for the three types considered.

AFINN

AFINN Python annotator calculates the scores taking into account each word. The score associated to each term, ranges between [-5;5], being -5 a very negative sentiment and +5 a very positive one. Nevertheless, for a given sentence, an accumulated value is retrieved. This value can be arbitrary low or high. Consequently, for the referenced variable, a MinMax normalisation is computed. Moreover, for scores lower than 0.5, the text is classified as negative. Scores equal to 0.5 refer to neutral texts and if these are greater than 0.5, are classified as positive.

Lexicon-of-abusive-words extended

For the case of lexicon-of-abusive-words extended version, each term from this dictionary has score values associated. These values quantify the level of hate contained in that word. Hence, for each text instance, an average score is computed by uniquely considering terms contained within the resource. Subsequent to apply a MinMax standardisation, a feature based in this lexicon is obtained. The higher the score value, the higher amount of abusive content has the instance.

The second step consists in calculating either the Mann-Whitney U test or the Wilcoxon Signed-Ranked test for quantitative data, as a subsequent step of checking whether observations are paired or not by applying the Spearman correlation test. The dependent variable is composed by quantitative features, meanwhile the independent variable is the label correspondent to each offense group, created from the variable of interest. The first quartile corresponds to the non offensive group and the fourth quartile to the highly offensive group. The statistical evaluation has the null hypothesis that variables extracted from taggers have the same mean value independently the offense group they belong to. The alternative hypothesis is that the means value differs, regarding the offense group it belongs to.

The results of this study can be observed in Table 3.10. The scores obtained by applying the linguistic resources for exploratory analysis show interesting results. Therefore, there are statistically significant differences between the two groups. However, in order to obtain a better interpretation, the meaning of each feature is briefly commented. *Valence* relies on characterising an emotion and describing if it is positive or negative, while *arousal* refers to the intensity of this emotion [7]. The *dominance* score consists in quantifying the feeling of an individual about the level of control related to their own personal life circumstances. A higher *dominance* value indicates that a person is decided, and has many ideas and opinions. Oppositely, lower values refer to an individual's feeling of being controlled/influenced by others. On the other hand, sentiment can be understood as a

Table 3.10: Tagger features in non and high offensive tweets in humorous subset.

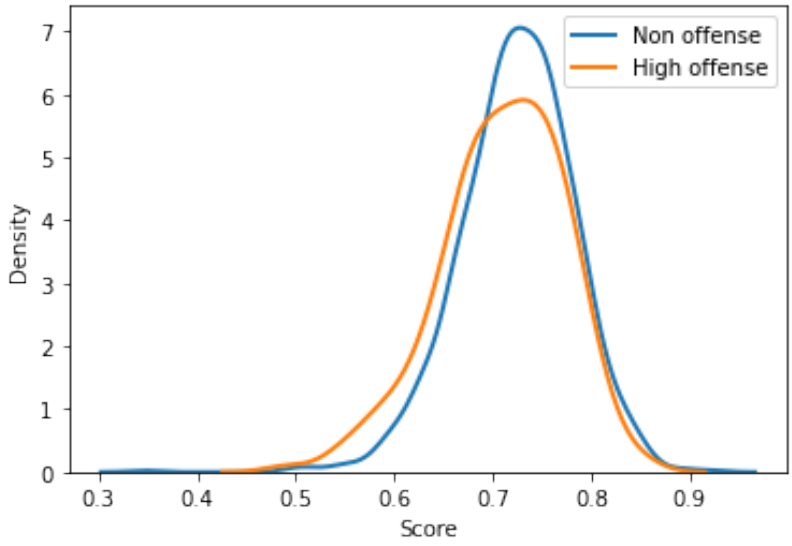
			Non offense Group 1		**High offense Group 4**	
Tool or Lexicon Name	**Feature**	**p-value**	**Mean**	**Variance**	**Mean**	**Variance**
SentiWordNet	Sentiment Score	0.0021	0.5188	0.0061	0.5084	0.0061
AFINN	Tagger - Valence Score	2.5e-11	0.6446	0.0041	0.6236	0.0049
VADER	Sentiment score	1.282e-11	0.0824	0.1908	-0.0476	0.1801
TextBlob	Polarity score	1.31e-07	0.0708	0.0758	0.0164	0.0621
	Subjectivity score	4.83e-07	0.4114	0.0995	0.35	0.0811
ANEW	Valence score	2.3e-10	5.7582	0.2126	5.6276	0.243
	Dominance score	4.23e-10	5.5608	0.0968	5.4778	0.1049
	Arousal score	0.011	4.0808	0.1116	4.1241	0.1504
Lexicon of abusive words extended	Score	0.003	0.4715	0.013	0.4836	0.0175

general positive/negative/neutral tone of the text. In the case of polarity and subjectivity scores, it is measured by the amount of opposite perspectives in a text and the amount of content based on personal beliefs, respectively.

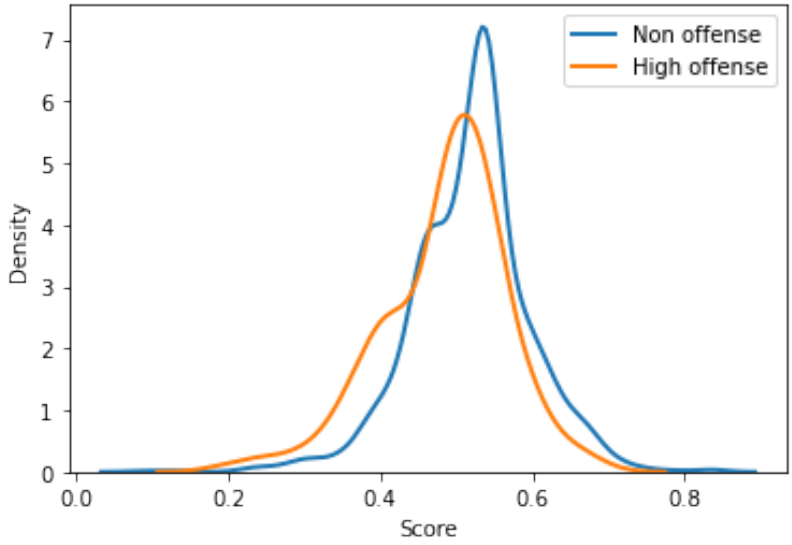
The sentiment score is larger in the non offensive set (M = 0.51) than in the high offensive one (M = 0.50) as shown in Figures 3.7a and 3.7b. The same pattern is observed for *valence* values. For non offensive instances, average *valence* scores are higher (M = 5.75 with $ANEW$) in the non-offensive set than in the high offensive one (M = 5.62 with $ANEW$), reflecting an overall positive emotion associated to the non-offensive humour. The difference in the use of abusive terms is also statistically significant (p-value = 0.003) and the average scores in the high offensive set (M = 0.48) is highest than in the non-offensive one (M = 0.47).

Differences observed between groups regarding polarity, display that the non-offensive set tends to contain a greater amount of opposite perspectives (M = 0.07) within the same text, than in the high offensive one (M = 0.01). In line with this result, subjectivity follows the same behaviour as polarity, noticed in Figure 3.7c: the non offensive set also shows a higher content based on personal beliefs (M = 0.41) than the high offensive one (M = 0.35). In this line, the level of *dominance* registered in the first group, exhibits a greater amount of powerful positions/points of view within the tweet's content (M = 5.56), than in offensive instances (M = 5.47). On the other hand, *arousal* is slightly lower in the non offensive set (M = 4.08) than in the high offensive one (M = 4.12). This result indicates that the content of this set is less likely to have stimulant/exciting words or terms linked to a strong emotional content, oppositely to the highly offensive set.

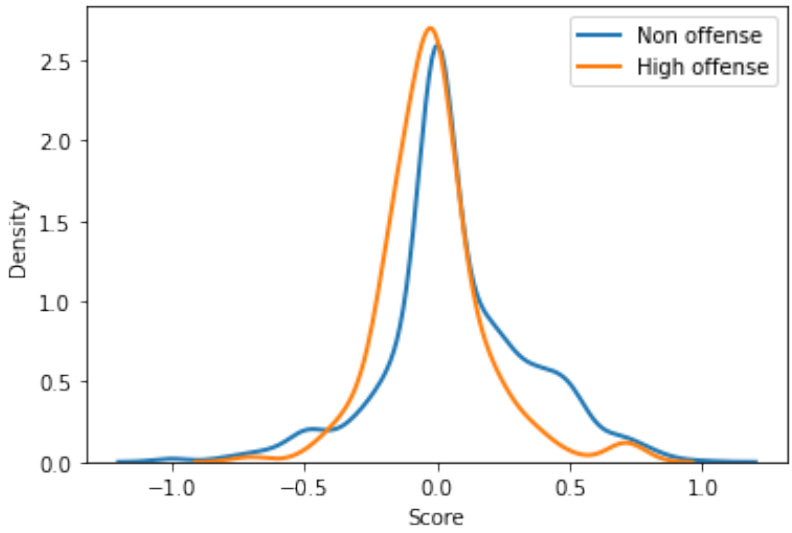
In summary, the offensive set of humorous texts has a greater amount of negative sentiment associated, in contrast to the non offensive one. Besides, texts in the offensive group show the same type of points of view. When comparing it with the non-offensive texts, the offensive ones tend to be less plural (have less different points of views), showing less subjective content than in the non offensive set. As a conclusion of this preliminary analysis, it can be confirmed that the manual



(a) Sentiment score - ANEW



(b) Sentiment score - SentiWordNet



(c) Polarity score - Texblob

Figure 3.7: Sentiment and polarity scores

annotation of the two groups of tweets present relevant differences between them by applying computational linguistic tools.

3.4.2 Features for Classification Task

Linguistic feature extraction conforms the core of this analysis. Hence, vectorized representation of features are achieved it through the Stanza tool and lexicons *Binary Lexicon-of-abusive-words, Hurtlex, EmoSenticNet, SentiSense and LIWC*, referenced as in Section 3.2. 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, it is 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. As a result, texts are represented as a frequency weighted term vector. Hence, each i-value of the linguistic feature corresponds to the ratio of occurrence of determined category inside the i-tweet.

3.5 Methodology

3.5.1 Feature Characterisation

Regularly, it remains necessary to provide a background to make decisions about diverse processes. With data as starting point, supplying context and extracting knowledge from series of either categorical and numerical variables is central to infer events. In order to develop experiments, a selection of non-parametrical statistical tests commented in Section 3.3 has been carried out. The principal assumption that most of the parametric statistical tests follow is that data have an unknown distribution, something observed in the features extracted, as shown in Figure 3.8.

Experiments applying these methods focus on analysing the impact it has over features the presence of humour and how the offense is reflected in jokes. By applying the commented methods, it has been possible to analyse linguistic features identified as statistically significant, taking into account their p-value and their presence within dataset. The first approach remains in detecting which are the linguistic features that characterise humorous texts. For this aim, 3,000 samples labelled as humorous and non humorous have been selected randomly, with a total of 6,000 observations, in order to carry out the statistical tests. Subsequently, within the humour set, a division by offense rating has been done by quartiles. Only considering outermost groups, the goal consists in detecting if linguistic features behave differently depending on the offense group of membership.

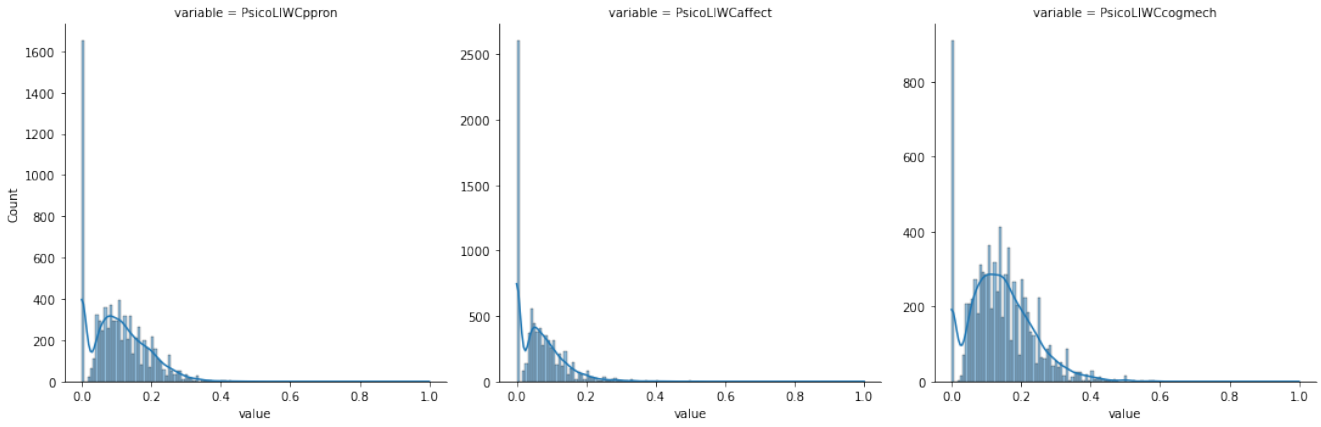


Figure 3.8: Features distribution

Firstly, the Spearman correlation has been computed in order to investigate if values of features were independent or not. If the null hypothesis is true, observations are not paired and the Mann-Whitney U test is used. If rejecting null hypothesis, it 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. Both, the Mann-Whitney U and Wilcoxon Signed-Ranked tests have been applied for testing differences between distributions of quantitative data. The independent variable has two levels: non offensive (1st quartile of "offense rating") and very offensive (4th quartile of "offense rating"). As a further step, the statistically significant features are classified in three groups regarding the characteristics of each one of them: affective, content and syntactic & morphological markers. The results of these experiments are presented in Chapter 4.

3.5.2 Classification Tasks and Ablation Test

For the second task, a binary classification is carried out. Besides seeking which variables characterise offensive humour, it is compelling to analyse their impact in their detection. Subsequent to identify which features are significant within the offensive humour characterisation, the most important ones are selected by considering their level of significance and their presence in the dataset.

Solely taking into account instances labelled as humorous, a division by quartiles regarding the "offense rating" feature is executed. As shown in Table 3.9, the first group corresponds to non offensive data, whereas group four contains the highly offensive instances. Hence, 2,420 observations composed by 1,210 samples from each group, are used in this task. Firstly, a baseline result is obtained, training models with all selected features. Afterwards, three extra experiments are computed under the form of ablation test. These experiments are executed by omitting in each iteration, syntactic & morphological, content and affective features. The objective of ablation test consists in detecting which set of variables represent a major impact within classifiers. Specifically, it enables to identify if there are specific features with higher relevance than others, by comparing the obtained results. The classifiers selected for this task are Support Vector Machine, Random Forests and

Logistic Regression, from Python library *scikit learn*[1]. Moreover, F_1-score is the selected measure for evaluating the systems performance. The results of these experiments are presented in Chapter 5.

To tune models, a grid search is carried out in order to optimise their hyperparameters. Support Vector Machine is trained with a linear kernel function, a regularisation parameter equal to 1,000 and a gamma of 0.001. Random Forest is trained with 30 estimators (number of trees in the forest), a maximum depth of the tree equal to 5 and 10 as the minimum number of samples necessary to split a node. Lastly, Logistic Regression is trained with a regularisation parameter equal to 1,000 and a "l2" penalty. The hyperparameters configuration above-mentioned is used in baseline classifier, as well as in all ablation tests.

[1]https://scikit-learn.org/stable/

Chapter 4

What Characterises a Good Joke?

4.1 Statistical Analysis

The experimental phase reported in this Chapter consists in two complementary approaches. Firstly, it is studied if the features behaviour differs in presence and absence of humour. In other words, analyse if exist and consequently identify, variables which characterise humour. "Is humour" is a binary variable, with 4,932 positive cases ("Is humour" = 1) and 3,068 negative ("Is humour" = 0) instances. For this aim, the data set was divided randomly into two subsets. In order to have balanced data, each set has 3,000 instances. One set is composed by texts labelled as humorous and the second one labelled as non humorous. The second approach, relies on observing which are the features that characterise offense within humorous tweets. That is, acknowledge if certain features can discern between presence and absence of offense being within jokes. Therefore, the total amount of humorous instances are applied in this study. Inside the humorous subset, only those instances annotated as humorous which belong to the first and fourth quartiles of "Offense rating" variable are included in the analysis. Specifically, the non offensive set corresponds to the first quartile, and is conformed by 1,253 instances. The high offensive set corresponds to the fourth quartile, composed by 1,210 examples.

In both steps, a seed has been fixed from Pandas[1] library, to make the results reproducible. For identifying which features are relevant in distinguishing humour from no humour, and offense/no offense within humour, several experiments have been carried out. For the purpose of inspecting which features are paired and which are not, a Spearman correlation test has been computed. Subsequently, the Mann-Whitney U test for independent data and the Wilcoxon Signed-Rank test for non independent data are applied. The null hypothesis consists in assuming that two random samples comes from the same population. Alternative hypothesis remains in assume that two random samples do not come from the same population. In the following sections, we compare the means of the two classes in several features. None of them presents a normal distribution. Therefore, non-parametrical tests are applied. For detecting which test applies, for each feature we compute the Spearman correlation between the two relevant classes (humour vs non humour or offensive vs non offensive humour). The aim is to detect whether samples are paired or not. If observing paired observations, the Wilcoxon Signed-Rank test is applied, otherwise, the Mann-Whitney U test is calculated.

[1] https://pandas.pydata.org/

4.2 Results

Linguistic features have been classified into the following groups, for better understanding and coherence in their analysis. Hence, features are characterised by specific markers.

Syntactic & morphological markers reflect the style of writing and the types of terms used. Taking this into account, these markers can be defined as the style of writing and elements of language that have been used in the text. Reflected through punctuation symbols and grammatical morphemes, these are elements which provide of coherence within texts [40] 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.

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 a lexical resource. A similar procedure is followed for features associated with emotions (anger, disgust, joy, like, love, sadness, surprise), which are terms associated with a person state.

Content markers indicate, as the name shows, terms related to content of sentence: concern words/terms within diverse categories likewise social groups, religion/sexual terms or hateful words, nature, and human-related terms.

4.2.1 Results for Humour Detection

Humour - No humour

Table 4.4 contains results of mean, variance and p-value of the most statistically significant features, with a confidence level of 95%. Same information for variables related to each group are shown in tables 4.1, 4.2 and 4.3. Complete results can be found in appendix A.

Table 4.1: Syntactic & Morphological markers - Significant features for humour detection

			Humour		**No Humour**	
Tool or Lexicon	**Feature**	**p-value**	**Mean**	**Variance**	**Mean**	**Variance**
LIWC	Personal Pronouns	1.27E-55	0.1162	0.0063	0.0847	0.0063
LIWC	Verb	2.08E-06	0.1475	0.0046	0.1412	0.0072
Part-of-Speech	Punctuation Symbol	7.50E-14	0.1119	0.0041	0.1061	0.0068
Part-of-Speech	Pronouns	1.36E-30	0.0673	0.0022	0.0547	0.0023
LIWC	I	7.45E-158	0.0587	0.0047	0.0198	0.0022
LIWC	Adverb	5.40E-05	0.0417	0.0022	0.0386	0.0024
LIWC	Article	3.00E-36	0.0805	0.0041	0.06	0.0033
LIWC	Prepositions	3.95E-32	0.0966	0.0038	0.1175	0.0052
LIWC	Auxiliary verb	0.00067	0.093	0.0031	0.0903	0.0045
Part-of-Speech	Adjectives	7.50E-05	0.0901	0.0046	0.0973	0.0052
LIWC	Past tense	2.98E-34	0.0364	0.0028	0.0201	0.0014

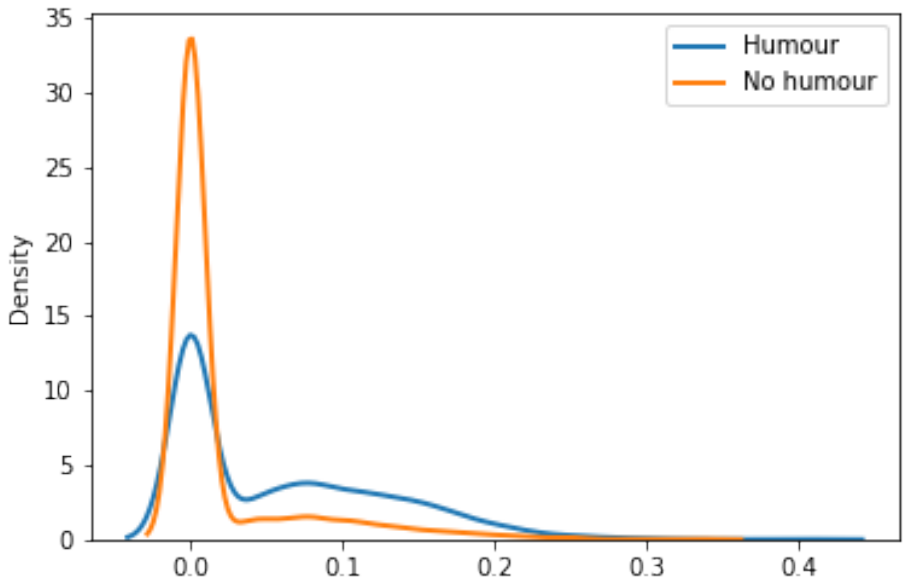


Figure 4.1: "I" distribution

Examining the set of features extracted in Table 4.1, it is noticeable that syntactic & morphological features remain as highly significant. Personal pronouns, pronouns in general and 1st person singular are mostly present in humorous instances (*"**I** like **my** coffee how **I** like **my** slaves... Free."*). As shown in Figure 4.1, distribution of "I" variable is right-skewed. The mean value in the non humorous set is centred in 0.0 with a peak in it, reflecting that majority of instances presents values with this score. In the humour case, the occurrence of "I" is way more variant. Moreover, the ratio of occurrence of humorous tweets, is higher in volume regarding no humour instances. A similar pattern is noticed for past tense (had, ran, ate), adverbs (sweetly, rapidly, well) and articles (the, a, an) although it is not so pronounced. In contrast, the 1st Plural Person has the opposite behaviour. The ratio of occurrence within the no humorous data is approximately triple regarding the humorous texts. Prepositions (of, to, at) and adjectives (beautiful, smooth, heavy) do not present marked disparities, but it still appears a statistically significant difference between the humorous and non humorous sets: prepositions and adjectives are more used in non humorous.

Regarding the content group, the most discriminant features for humour are related to biological processes and humans (sexual, male genitalia, body, biology, hear and family). Besides presenting a higher occurrence within humour, some terms connected to distinct categories are related in the same tweet, e.g. *His **son** asked him what **gay** meant. **Son**: **Dad**, what does **gay** mean? **Dad**: Happy **son**. It means happy. **Son**: Then are YOU **gay** **DAD**? **Dad**: No **son**..... i have a **wife**...* On the other hand, it is relevant to remark that features related to abusive words score show a significant difference between both groups. A higher score corresponds to a greater presence of abusive words within tweets, while lower scores are associated to less abusive terms. Therefore, as observed in Table 4.2, the humorous set presents a higher mean score for features associated to abusive terms (M = 0.0013 with Hurtlex and M = 0.001 with the Abusive words resource), in comparison to no humour (M = 0.0007 with Hurtlex and M = 0.0007 with the Abusive words). This result is effective to confirm the fact that humour is used mostly to hurt in comparison to no humour.

Table 4.2: Content markers - Significant features for humour detection

			Humour		No Humour	
Tool or Lexicon	**Feature**	**p-value**	**Mean**	**Variance**	**Mean**	**Variance**
LIWC	Cognitive Mechanism	4.42E-41	0.1267	0.007	0.1619	0.0101
LIWC	Relativity	5.06E-06	0.1187	0.0078	0.1314	0.0095
LIWC	Biology	0.0052	0.0464	0.0034	0.0442	0.0042
Hurtlex	Male Genitalia	5.62E-32	0.0287	0.0016	0.0179	0.0011
LIWC	Quantifiers	4.10E-16	0.0222	0.0013	0.0302	0.0018
LIWC	Achieve	4.70E-40	0.0168	0.001	0.0305	0.0022
LIWC	Body	1.31E-05	0.0164	0.0012	0.0134	0.0011
LIWC	Insight	8.05E-30	0.0147	0.0008	0.0251	0.0016
LIWC	Hear	4.41E-47	0.013	0.0008	0.0053	0.0005
LIWC	Health	1.50E-05	0.0124	0.0009	0.0162	0.0013
LIWC	Sexual	7.15E-08	0.0114	0.0009	0.0079	0.0007
LIWC	Family	6.17E-38	0.0107	0.0007	0.0039	0.0003
Hurtlex	Negative Stereotypes & Ethnic Slurs	1.53E-22	0.0043	0.0003	0.0007	3.39e-05
Abusive Words	Binary Lexicon	0.00010	0.0013	7.64e-05	0.0007	6e-05

Furthermore, features with terms related to family are widely present in humorous texts (M = 0.0107), rather than in non humorous ones (M = 0.003), as shown in Figure 4.2 and Table 4.2. Besides, one hypothesis about humour is that most of jokes have negative connotations, and also, a good part of them involves a familiar context, sometimes making use of several word/terms from distinct categories, to create jokes against certain social groups. For example: *A **gypsy girl** tells her **mum** she's **pregnant**, "Congratulations" says her **mum**... "Do you know who the **father** is?"... "**Mum**.... if you **ate** a tin of beans would you know which one made you **far***?"*.
Moreover, ratio of terms linked to achievement (work, study, hero), cognitive processes (cause, know, must) and insight (think, meditate, consider) are more present in non humorous tweets that in the humour class. For example: *Better to **try** and **fail** at something **important**, than to **succeed** in something that isn't*, and, ***Knowledge** is not enough; we **must apply**. **Intention** is not enough; we **must do**..*

Tackling affective variables (see Table 4.3), the results confirm their relevance on humour detection. It is important to notice that the emotions which differentiate humour from no humour, are mostly negative (surprise, disgust, anger). For example: *What's the difference between a **shi*** golfer and a **shi*** skydiver? The **shi*** golfer goes: Wham! Damn! The **shi*** skydiver goes: Damn! Wham!.*
A lower occurrence of positive emotions, noticed when comparing (observed in Figure 4.3) humorous texts (M = 0.02) and no humorous ones (M = 0.04), also confirms that many times humour is used to hurt. Same pattern are observed in the expressions of joy, like and love. Furthermore, instances labelled as no humorous, with a higher amount of these terms, tend to be mostly personal related (*Send her a **good morning** text and she'll **love you** forever.*) or linked to quotes (*Your **toughest challenges** have the **best rewards***). Although the humour set also has other emotions in it (antici-

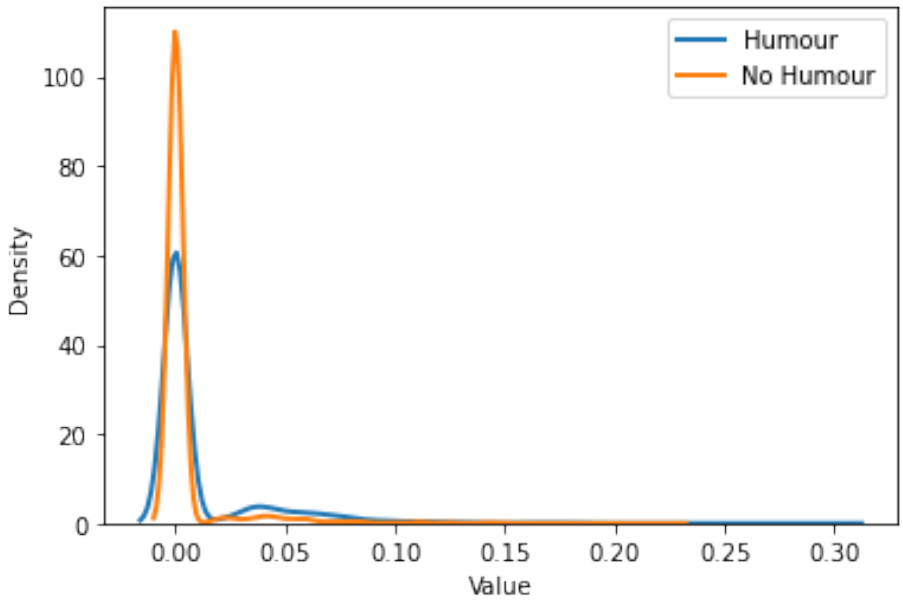


Figure 4.2: Family distribution

Table 4.3: Affective markers - Significant features for humour detection

			Humour		**No Humour**	
Tool or Lexicon	**Feature**	**p-value**	**Mean**	**Variance**	**Mean**	**Variance**
EmoSenticNet	Joy	2.25E-30	0.3125	0.013	0.3481	0.0138
LIWC	Affective Processes	5.00E-31	0.0564	0.0037	0.0779	0.0061
EmoSenticNet	Surprise	4.80E-21	0.0482	0.0041	0.0297	0.0017
SentiSense	Disgust	0.00028	0.032	0.0018	0.0274	0.0016
LIWC	Positive Emotions	1.18E-36	0.0293	0.0019	0.0486	0.0043
SentiSense	Like	5.47E-18	0.0261	0.0015	0.0362	0.0025
SentiSense	Anticipation	2.08E-15	0.0124	0.0007	0.0184	0.0011
EmoSenticNet	Fear	0.0002	0.0116	0.0008	0.0144	0.0009
LIWC	Anger	1.10E-08	0.0114	0.0007	0.0082	0.0007
LIWC	Sad	1.22E-10	0.006	0.0004	0.0097	0.0008
SentiSense	Love	1.23E-13	0.0051	0.0003	0.0096	0.0007
LIWC	Inhibition	1.08E-13	0.005	0.0003	0.0085	0.0006
LIWC	Anxiety	9.41E-07	0.003	0.0002	0.0054	0.0004

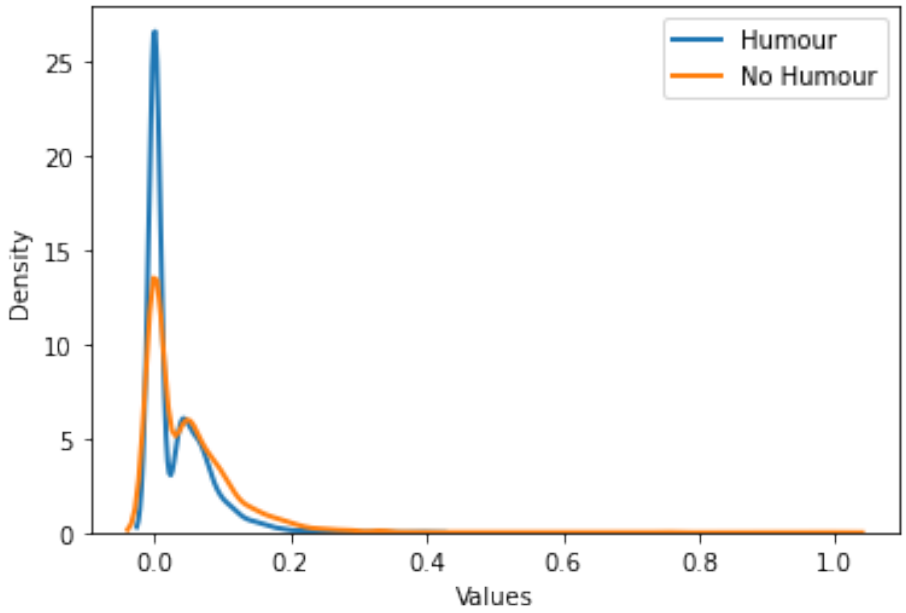


Figure 4.3: Positive emotions distribution

pation, fear, sad, anxiety), their quantity is the half when comparing with the opposite set (*"**Broke, depressed**, and **emotionally vacant**." - Macklemore.*). In line with the previous conjecture, it can be added that negative emotions associated to mental health issues tend to appear in no humorous texts. Meanwhile, emotions mostly provoked by third-parties remain present in humorous tweets.

Table 4.4: Most significant features for humour detection

			Humour		No Humour	
Tool or Lexicon	**Feature**	**p-value**	**Mean**	**Variance**	**Mean**	**Variance**
LIWC	I	7.46E-158	0.0587	0.0047	0.0198	0.0022
Part-of-Speech	1st Person Singular	5.95E-119	0.0318	0.002	0.0111	0.001
LIWC	Plural Pronouns	1.27E-55	0.1162	0.0063	0.0847	0.0063
LIWC	Pronouns	1.43E-49	0.1669	0.0071	0.1342	0.0097
LIWC	Hear	4.42E-47	0.013	0.0008	0.0053	0.0005
LIWC	Cognitive Mechanism	4.43E-41	0.1267	0.007	0.1619	0.0101
LIWC	She/He	6.53E-41	0.0185	0.0014	0.0098	0.0011
LIWC	Achieve	4.70E-40	0.0168	0.001	0.0305	0.0022
LIWC	Family	6.17E-38	0.0107	0.0007	0.0039	0.0003
LIWC	Positive Emotions	1.19E-36	0.0293	0.0019	0.0486	0.0043
LIWC	Article	2.99E-36	0.0805	0.0041	0.06	0.0033
LIWC	Past tense	2.98E-34	0.0364	0.0028	0.0201	0.0014
LIWC	Prepositions	3.95E-32	0.0966	0.0038	0.1175	0.0052
Hurtlex	Male Genitalia	5.63E-32	0.0287	0.0016	0.0179	0.0011
LIWC	Affective Processes	5.03E-31	0.0564	0.0037	0.0779	0.0061
Part-of-Speech	Pronouns	1.36E-30	0.0673	0.0022	0.0547	0.0023
EmoSenticNet	Joy	2.25E-30	0.3125	0.013	0.3481	0.0138
LIWC	Insight	8.06E-30	0.0147	0.0008	0.0251	0.0016
LIWC	We	8.07E-29	0.0042	0.0003	0.0127	0.0012
LIWC	Swear Words	1.39E-25	0.005	0.0004	0.0014	0.0001
Hurtlex	Negative Stereotypes and Ethnic Slurs	1.53E-22	0.0043	0.0003	0.0007	3.39e-05
EmoSenticNet	Surprise	4.81E-21	0.0482	0.0041	0.0297	0.0017
Part-of-Speech	1st Plural Person	2.65E-20	0.0022	0.0001	0.0065	0.0005
SentiSense	Like	5.47E-18	0.0261	0.0015	0.0362	0.0025
LIWC	Quantifiers	4.16E-16	0.0222	0.0013	0.0302	0.0018
SentiSense	Anticipation	2.08E-15	0.0124	0.0007	0.0184	0.0011
LIWC	Certainty	2.68E-14	0.0113	0.0006	0.0182	0.0013
Part-of-Speech	Punctuation Symbol	7.53E-14	0.1119	0.0041	0.1061	0.0068
LIWC	Inhibition	1.09E-13	0.005	0.0003	0.0085	0.0006
SentiSense	Love	1.23E-13	0.0051	0.0003	0.0096	0.0007

4.2.2 Results for Offense Detection in Humour

Offense - Non offense

Similarly to the humour detection analysis, the results for offense detection are provided in Table 4.8. It includes the values of mean, variance and p-value of those features which present a statistically significant difference, with a confidence level of 95%, between the two groups of non-offensive jokes and offensive ones. As in humour analysis, relevant features are classified into syntactic & morphological, affective and content groups. Their results locate in Tables 4.5, 4.6 and 4.7 respectively. Complete results are in appendix B.

Syntactic & Morphological features

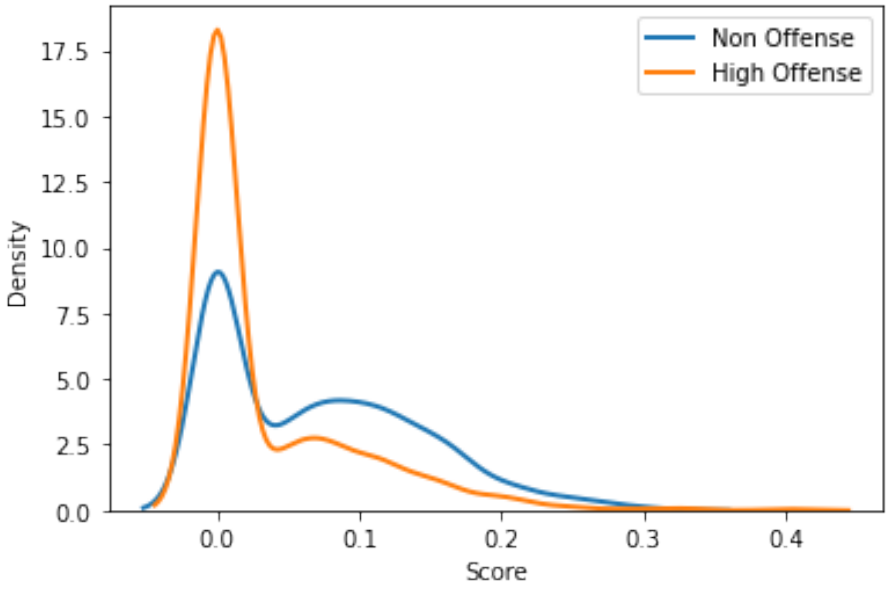
The results observed in Table 4.5, can be considered as relevant for offense detection. The behaviour of personal pronouns, first person singular/plural and second person singular matches in both scenarios. Their ratio of occurrence is higher in non offensive tweets than in offensive ones. However, the third person of plural follows an opposite pattern. Both behaviours can be observed in Figure 4.4. Hence, hurtful/hate speech contained in jokes tends to be intended to minorities and not to a specific individual. 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 the words he/she says and removes any possible guilt.

Although being highly present in offense and no offense 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 applied in offensive contexts. As articles define a noun as specific or unspecific, this use of the articles, 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 are likely 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.*

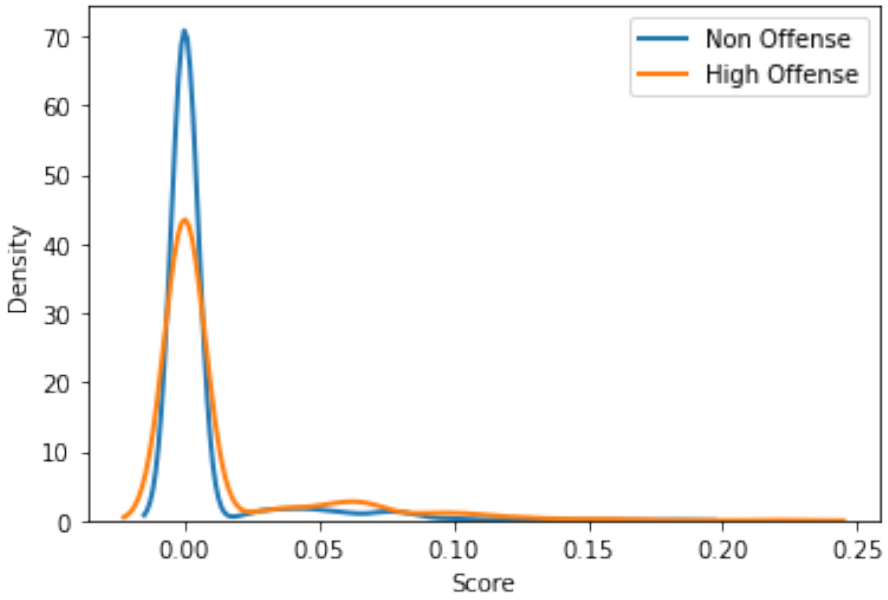
Affective features

In terms of emotions, results coincide regarding humour analysis while being relevant for offense distinction. Negative emotions (anger, disgust, fear and sadness) appear to be highly present through offensive tweets than in the non offensive ones. Moreover, the offensive set presents a high amount of terms related to surprise, an emotion that could be positive or negative, as shown in Table 4.6 and Figure 4.5. Affective processes and positive emotions in general tend to appear mostly in healthy jokes.

A different trend is visible for joy terms. Table 4.6 reports 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 *gay* term, is associated with this emotion no matter if is a term associated



(a) I distribution



(b) They distribution

Figure 4.4: Differences in “I” and “They” between offensive and non offensive sets

Table 4.5: Syntactic & Morphological features belonging to non and high offense rating tweets in humorous subset.

			Non Offense		High Offense	
Tool or Lexicon Name	**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
Part-Of-Speech	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
Part-Of-Speech	1st Plural Person	3.25E-06	0.0033	0.0002	0.0012	0.0001
Part-Of-Speech	Adverbs	2.91E-05	0.0566	0.0033	0.048	0.0031
Part-Of-Speech	Noun	8.87E-05	0.2511	0.0088	0.2379	0.0092
Part-Of-Speech	2nd Person Singular	4.93E-03	0.0013	0.0001	0.0005	3.0e-05

to a sexual orientation. The word gay dates back to the 12th century and comes from the Old French "gai," meaning "full of joy or mirth.". That could be the reason why *SentiSense* associates this term among the ones related to the joy emotion: *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*.*

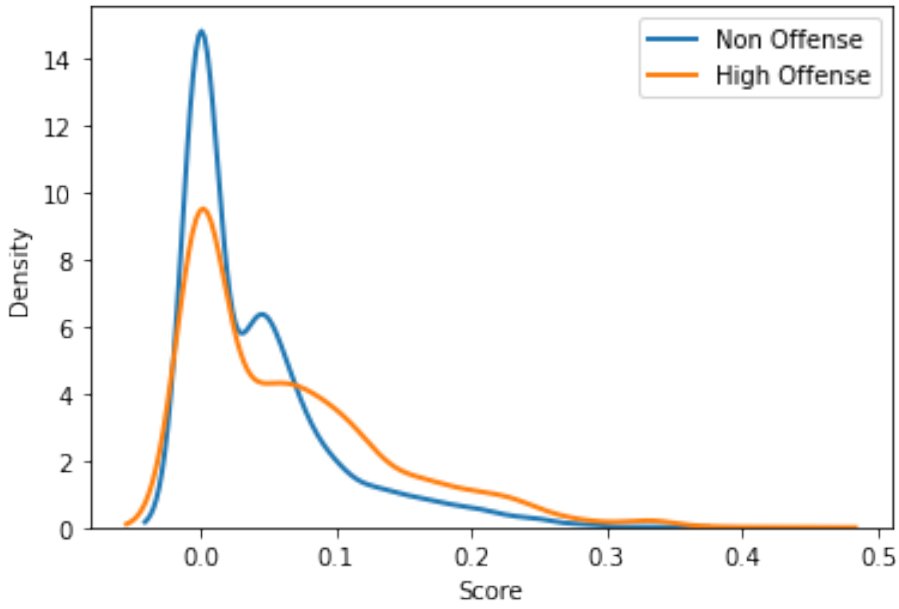


Figure 4.5: Surprise distribution

Table 4.6: Affective features belonging to non and high offense ratings tweets in the humorous subset.

			Non Offense		High Offense	
Tool or Lexicon Name	**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

Content features

Results regarding the content group are observed in Table 4.7. Content features relate the overall topic of the studied texts. In this case, it is possible to infer that majority of tweets contained in this particular data subset, are about human race and social terms. It is noticeable that words associated to referenced topics (biology, humans, sexual, see, social, religion, negative stereotypes and ethnic slurs, moral and behavioural defects, swear words) are mostly used (see Figure 4.6) when they come from hurtful jokes than in non offensive ones, likewise: *Where do most **black people** work? In **jail**.*

The most notorious differences between non offensive and offensive texts are observed (see Table 4.7) in features with jokes regarding sexuality (gay, lesbian, prostitute), religion (Jewish, christian, Christmas), swear words (*Don't blame **Christmas**. You were **fuc***ng fat** in August.*), negative stereotypes and ethnic slurs (Mexican, Chinese, black people) (see Figure 4.7) and moral & behavioural defects (jail, death). On the other hand, variables about leisure (canoe, cook, chat) and exclusive (but, without, just) are the only ones which are less frequent in offensive texts (***Bathtub** is **just** a reverse **canoe**.*).

In summary, in Chapter 4 it has been observed that some characteristics that differentiate humorous from non humorous texts, also serve to differentiate between non-offensive and offensive humour. This is the case, for example, of the presence of negative emotions or words referring to negative stereotypes, ethnic minorities or sexual content. The same occurs with the personal pronoun "they", which is more present in highly offensive humour than in non offensive jokes. Hence, these results suggest the proximity between offensive humour and hate speech.

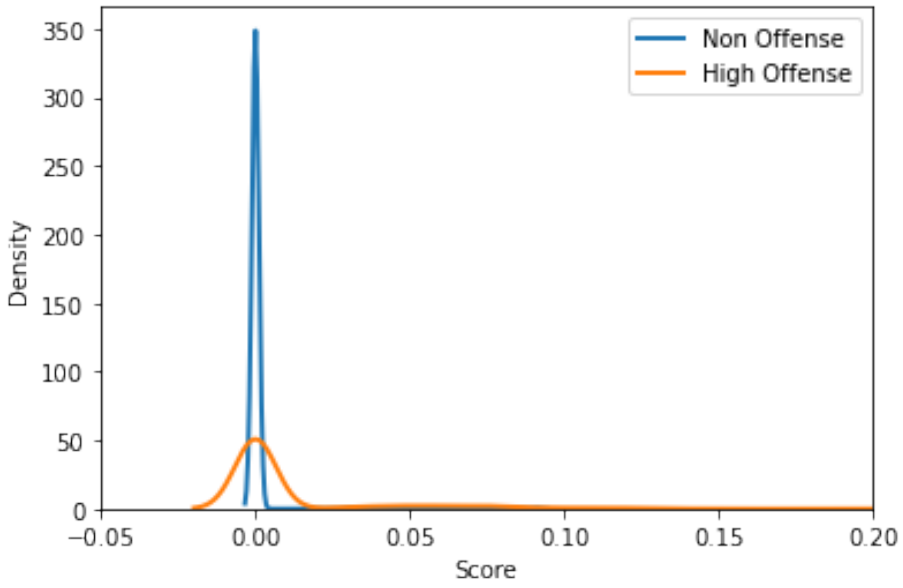


Figure 4.6: Negative stereotypes and ethnic slurs distribution

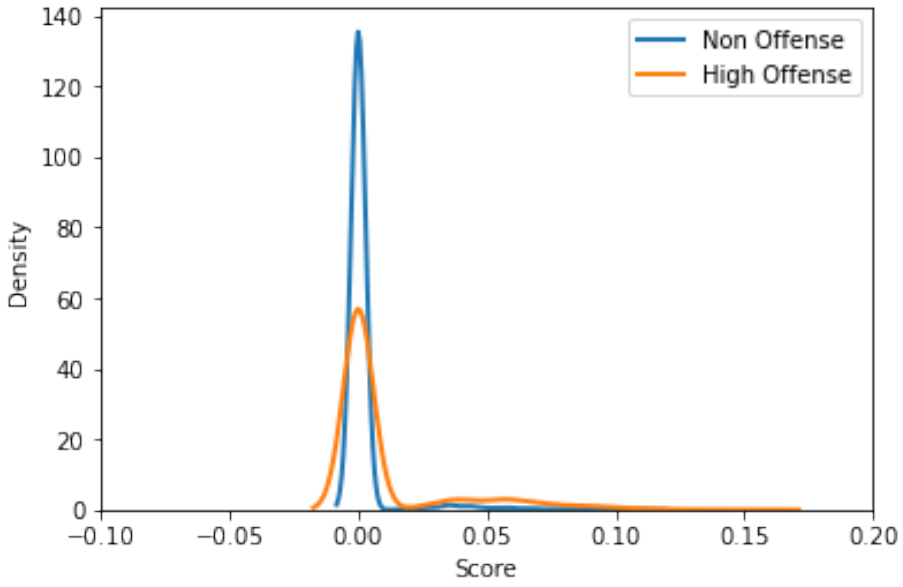


Figure 4.7: Moral & behavioural defects distribution

Table 4.7: Content features belonging to non and high offense ratings tweets in the humorous subset.

			Non Offense		High Offense	
Tool or Lexicon Name	**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 and 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

After identifying the significant groups of variables that differentiate non offensive humour from highly offensive humour, a step further is taken. In order to study how relevant variables behave in offense detection inside humour, a classification task is carried out. It is important to remark that some statistically significant features provide the same information, given that they have been extracted from distinct resources (e.g. anger extracted from *SentiSense*, *EmoSenticNet* and *LIWC*). As a consequence, in order to avoid redundant data, the most significant ones (see table 4.8) are selected to be used by classifiers.

Table 4.8: Features used by the classification system.

			Non Offence		High Offence	
Tool or Lexicon Name	**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
Part-Of-Speech	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
Part-Of-Speech	1st Person Plural	3.25E-06	0.0033	0.0002	0.0012	0.0001
Part-Of-Speech	Adverbs	2.91E-05	0.0566	0.0033	0.048	0.0031
Part-Of-Speech	Noun	8.87E-05	0.2511	0.0088	0.2379	0.0092
Part-Of-Speech	2nd Person Singular	0.005	0.0013	9.0e-05	0.0005	3.0e-05
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
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 and 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

Chapter 5

Classification of Non and Highly Offensive Humour

5.1 Introduction

This part of the study puts the focus on the classification of the grade of offense within humour and an analysis on the contribution of affective, syntactic, morphological and content features. To perform the study, only instances annotated as humorous which belong to first (non offensive) and fourth (highly offensive) quartile of offense rating scores are included. The criterion of feature selection for this task is composed by considering the most significant variables within the three categories of features: affective, syntactic & morphological and content features, as referenced in Section 4.2.2. The execution of experiments is performed by dividing the data in 80% training set and 20% testing set. 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 classification systems applied are: Support Vector Machine (SVM), Random Forests (RF) and Logistic Regression (LR). For evaluating the performance of the models, it is computed on testing set several measures: accuracy, recall, precision and F_1-score. Each measure is briefly explained, in order to provide a better interpretation of the classifiers results.
Accuracy [17] consists in the ratio of predictions correctly classified. Recall (also known as True Positive Rate), gives answer to the following question: Which percentage of true positive cases are detected? The recall score [17] refers to the proportion of known positives that are predicted correctly. Both TP (True Positives) and FN (False Negatives) are applied in the calculus of this measure (Figure 5.1). Precision (also known as Positive Predictive Value), gives answer to the question: Which percentage of cases predicted as positive are correct? Alternatively stated, the precision [17] measures the proportion of cases predicted as positive which are truly positive. Both TP (True Positives) and FP (False Positives) are applied in the calculus of precision, as shown in Figure 5.1. As an aggregate metric, F_1-score [17] is one of the most common classification measures. It is calculated as the harmonic mean of the precision and recall, as shown in the next expression:

$$F_1 = 2 * \frac{precision * recall}{precision + recall}$$

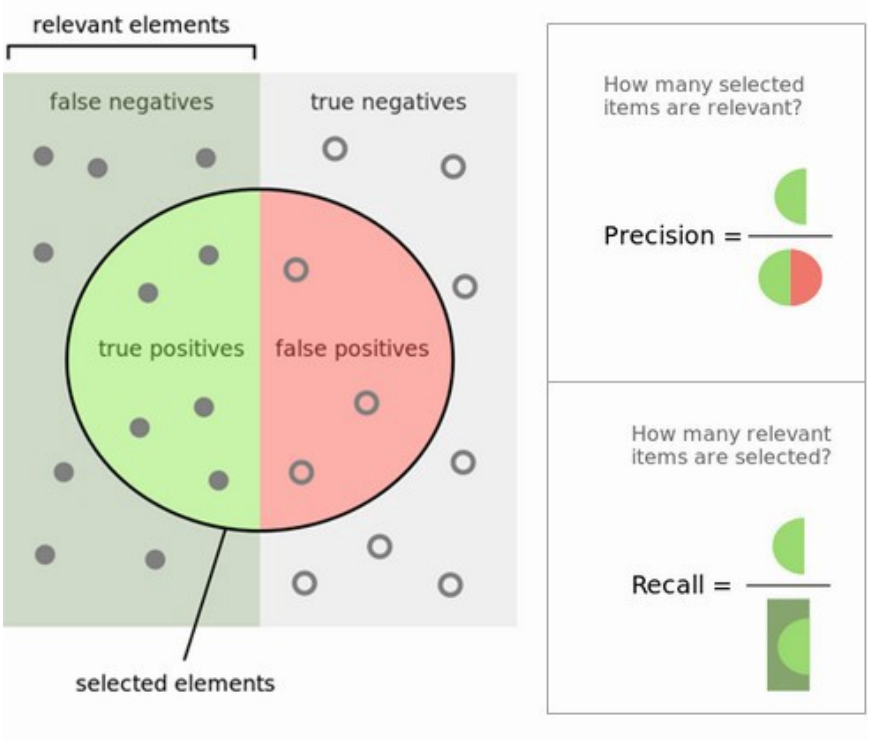


Figure 5.1: Precision and recall measures [2]

5.2 Classification Task

The classifiers are trained with the 35 most relevant linguistic features (see Table 4.8). Table 5.1 provides the results obtained by each classifier. All models perform better when classifying the non offensive class, achieving a greater F_1-score than for the highly offensive class. Although differences among scores are not extremely high in any model, Support Vector Machine is revealed as the one which faces greater difficulties in class differentiation. However, F_1-score Macro ranges in between 73%-76%, showing that models present a similar performance. The same pattern is observed in terms of accuracy results, being the RF model the most accurate one with an accuracy and a F_1-score of 76% (Table 5.1), followed by the SVM and the LR classifiers. Table 5.3 reports the resulting confusion matrix of this approach. In general, observing the percentage of true negative cases, it confirms that good results are obtained for non offensive cases. Moreover, it is noticeable how the positive class is misclassified, showing a poor performance overall. Nevertheless this situation is expected, considering the difficulty of models to detect offensive instances, as previously commented.

Furthermore, it is observed a good precision for the RF model, given the low value of FP (False Positive) rate and high TP (True Positive) rate obtained by it. This result can be interpreted as a better performance of this classifier for positive class detection. Moreover, the Random Forest model is the one which present the highest recall ratio. Values of FN (False Negative) and TP stands out from the ones obtained by SVM and LR. This reflects that the RF model detects a high number of truly positive instances. Therefore, it is observed that the classifier constructed with Random Forest is one step ahead from LR and SVM systems.

Table 5.1: F1-score & Accuracy in the classification task

	Non offense	**High offense**		
Model	**F1-score**	**F1-score**	**F1-score Macro**	**Accuracy**
Support Vector Machine	0.76	0.72	0.74	0.74
Random Forests	0.77	0.75	0.76	0.76
Logistic Regression	0.74	0.72	0.73	0.73

5.3 Ablation Test

In order to quantify the contribution provided by each group of features in the classifiers performance, an ablation test has been done. Each classifier has been trained with different combinations of groups of features (affective and content, content and syntactic & morphological, affective and syntactic & morphological). With the groups related either to affective, syntactic & morphological, and content features, it is observed which generates a stronger decrease on model performance and which one does not produce remarkable changes. Table 5.2 shows that, in general, all systems perform worse when removing any group of features, while Table 5.3 reflects how ablation study shows the systems capability of discriminating between classes.

The effect that the affective group has over models performance, are noticed in the values of TP and FN. Specifically, it is observed a deterioration on the performance of SVM and RF models. Furthermore, Random Forest is the classifier which has the strongest decrease in these values, being also reflected in the worsening of the recall measure (5% decrease), as shown in Table 5.3 and in the value of the F_1-score (4% decrease) shown in Table 5.2. The unique exception occurs for Logistic Regression when removing affective features. It is observed a subtle increase in F_1 metric (Table 5.2).

Table 5.2: Ablation test for SVM, RF and LR

	Support Vector Machine	**Random Forest**	**Logistic Regression**
	F1-Score Macro	**F1-Score Macro**	**F1-Score 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)

Looking at FP and TN ratios (Table 5.3), there is a slight improvement in their values. However, it is noticeable how the decrease of misclassified instances in the positive class (lower FP) widely contributes to the increase of 3% in the precision score in comparison with the LR model trained with all features (Table 5.3). This result could be explored in more detail as a future work.

The syntactic & morphological group produces a similar pattern of models behaviour regarding F_1-score. The removal of this group generates a decrease in F_1 metric for all the classifiers (Table 5.2). The systems that do not contemplate this group, show a greater percentage of offensive instances misclassified (increase in the FN ratio) and less capability of classify positive cases properly (decrease in the TP ratio), as shown in Table 5.3. Meanwhile, FP and TN values improve in SVM and LR, contrary to RF which does not present any variation in these ratios. Changes in TP and FN ratios can also be clearly noticed on the recall percentages observed in Table 5.3. That is to say, removal of syntactic & morphological group decreases for all the models, the proportion of well-captured cases known as positives. On the other hand, the changes in the FP and TP ratios can be noticed on the precision percentages (Table 5.3), however, this changes are not as pronounced as the ones observed in the recall values. Although the fact that for the RF model, the precision value subtly decreases, for SVM it remains as before, while improving for the LR system. That is to say, the removal of this group do not change in a notorious way, the capability of predict positive cases which are truly positive. Nonetheless, it is clear that the behave pattern between the classifiers vary.

Content group remains as the most relevant set of features. By removing this set, an important reduction in the values of overall metrics is observed. Although F_1 metric persists as competitive (0.65 - 0.66), it decreases in almost a 10% from the F_1-scores observed for the models trained with all the features. Furthermore, there is a notorious difference in the F_1 measure obtained with previous approaches, as shown in Table 5.2. This goes in consonance with FP, FN, TP and TN rates displayed in Table 5.3. These values worsened when removing this set of features, strongly increasing the FP rate and decreasing the TN rate. As a consequence, the proportion of known positives that are predicted correctly (recall) is deeply altered. Although precision percentages also show a decrease in their values, it is not as pronounces as in the case of the recall percentage. As a result, regarding content features, these are identified as the ones which contributes in a substantial manner, for all systems.

5.4 Discussion

As commented before, offensive language is present in a wide range of jokes, covering a long set of sensitive tasks. A few of them are related to religion, sexual terms, morality, and human-related terms. Given that offense is frequently disguised as humour within jokes, people sometimes face issues in identifying it, as well as classification systems. Classifiers trained with the three groups of features distinguish better “Non offense”, or what is the same, face difficulties in offense detection. Nevertheless, slight differences are observed between models. Random Forest is yielded as the best in terms of offense detection, with a F_1-score equal to 75% and a F_1 Macro of 76% (Table 5.1).

In line with the previous section, more important differences are detected among models in ablation study. Firstly, the syntactic & morphological group helps distinguish better “Non offense” class in the SVM and LR models. Moreover, it affects badly all over in offense detection. Considering LR, the affective group helps obtaining a good classification in both classes. Furthermore, it also helps the SVM system to better detect non offensive instances. Tacking into account the absence of content features, these contributes negatively for detecting either “Non offense” and “Offense”

Table 5.3: Values of confusion matrix for the models with all features and for ablation test

		False Positive FP(%)	**False Negative FN(%)**	**True Positive TP(%)**	**True Negative TN(%)**	**Recall(%)**	**Precision(%)**
All features							
	Support Vector Machine	8	18	34	40	66	81
	Random Forests	8	16	35	41	69	82
	Logistic Regression	11	16	35	38	68	77
Affective							
	Support Vector Machine	8	19	33	41	64	81
	Random Forests	9	18	33	39	64	78
	Logistic Regression	9	16	35	40	68	80
Syntactic & Morphological							
	Support Vector Machine	7	20	31	41	61	81
	Random Forests	8	19	33	41	63	80
	Logistic Regression	10	18	33	39	64	78
Content							
	Support Vector Machine	17	19	33	32	63	66
	Random Forests	16	18	34	33	65	68
	Logistic Regression	16	18	34	32	65	67

instances. However, it is important to remark that the decrease on the detection of non offensive tweets is more pronounced than the decrease on the detection of offensive ones, when removing content features from the classifiers training.

It is interesting to note that the performance of Random Forest is the most sensitive to the absence of affective, content and syntactic & morphological groups, reflected in the F_1 values shown in Table 5.2. Although the SVM follows the same pattern of RF when it comes to performance, is not as sensitive as RF, in terms of F_1-score. Meanwhile, Logistic Regression improves without affective markers, whereas the remaining groups produces a decrease on the LR performance.

Chapter 6

Conclusion and Future Work

6.1 Conclusion

This Chapter summarises the main results of the research carried out, as well as discuss future work. The main results are summarised attending to the research questions.

RQ1. Which are the features that distinguish humorous texts from the non humorous ones?
As an introductory way to analyse humour and non humour, it is considered as relevant to study which features discriminate better between both of them. Regarding affective terms, it has been observed that jokes have an amount of negative emotions greater than non humorous text. Meanwhile, non humorous texts are the ones which present mostly positive words (relative to love, joy, like and positive emotions in general). Variables related to personal pronouns, use of prepositions, human-related terms (biology, family, sexuality, abusive terms, stereotypes and ethnic slurs) are widely present in humorous instances. The fact that the lexicon of dictionaries such as *Hurtlex* or *Abusive words* is more present in humorous texts than in non-humorous ones, confirms the starting point of this research: humour is used in many cases to hurt. Also it is important to notice that the offensive set of humorous texts are less plural (have less different points of views), showing less subjective content than in non offensive set.

RQ2. Which are the features that distinguish non offensive humour from the offensive one?
The use of terms related to personal pronouns, and mostly to first singular person "I", changes if the humour is offensive or not. Furthermore, the use of these types of words are minor in offensive instances. Hence, the authors of offensive jokes tend to do it from a depersonalization position. Moreover, it has been detected that offensive tweets do contain a high number of terms related to nouns, articles, adjectives and third plural person pronouns. Linked to the last feature, the target of these jokes tend to be groups of people, rather than an individual.
Results with affective markers confirm the initial intuition. In fact, terms related to negative emotions are mostly present in offensive instances, while positive ones are widely represented in non offensive jokes. As expected, offensive tweets are closely related to anger, disgust, sadness, fear and surprise. Therefore, these results go in hand with previous one, showing a great amount of negative

connotations against groups within these jokes. In the last place, the features linked to human body, ethnics, religion, defects in general and sexual content help to differentiate better offensive jokes from non offensive ones. However, non offensive humorous texts tend to have a greater amount of terms related to *see* and *exclusive words* (but, instead, without), unlike offensive jokes.

RQ3. How do classifiers perform distinguishing non offensive humour from the offensive one? It has been observed that classifiers obtain a better performance on non offensive humour, while offensive jokes have a high percentage of misclassification. As mentioned previously, the offense relation with humour rating is inversely proportional. Hence, it is possible that these features are linked in a subtle way, likewise implicit abuse, producing a difficulty in its detection. Regarding model performance, Random Forest presents higher values of accuracy when trained with all features. Moreover, Random Forest is the system which has the highest F_1 measure in the baseline, nevertheless, it is the most sensitive model when analysing it results within ablation test.

RQ4. Which are the characteristics that enable classifiers to make this distinction? On the other hand, linguistic features yielded interesting results related to their contribution to the models. Content features are the ones which contribute the most to models performance. This means that for the discrimination of offensive and non offensive humour, the topic of jokes is the most important aspect to consider. Offensive jokes have a high content of terms related to sexuality, biology, religion, negative stereotypes, ethnic slurs, moral and behavioural defects and swear words. Syntactic & morphological features also help to differentiate between the two types of subsets, as these variables stress the jokes target, as marked in personal pronouns behaviour. Moreover, with affective features it is possible to observe the emotional content of both groups. Furthermore, these features shows how offensive tweets present the greatest amount of terms related to negative emotions, unlike non offensive ones.

Our results go in line with the ones obtained in studies such as [22] and [18]. These reveal, from a psychological point of view, that humour is a method to express ideas related to topics considered taboo, racist or that have a negative connotations in general. By expressing these preconceived ideas in comic contexts, people tend to feel less responsible/guilty about the opinions told by them. Furthermore, as it has been detected a high amount of words related to third plural person, it can be inferred that targets of jokes are composed by groups. In addition, as commented before, the low presence of words related to the first person can confirm the fact that the authors of jokes tend to untie about anything they said.

6.2 Personal Assessment

From a personal point of view, this project has contributed to me both personally and professionally. On the professional side, I could improve my programming skills, owing to data preprocessing, linguistic feature engineering, and models development. Moreover, I learnt about non-parametrical

statistical methods that composed a relevant task in this project. Applying several lexical resources, lexical Python packages, learning new statistical methods (e.g. non parametric ones) and *scikitlearn* for modelling, as well as research on topics related to psychological, are some of the acquired knowledge. However, the most remarkable acquired technical knowledge is related to human behaviour. Learning how comics statements can turn into hurtful ones, showed how much witty people can be, besides stressing out currently present social differences within society.
On the personal side, time management has been fundamental for the development of this final project. However, the improvement of my communication skills required for providing an attractive story-telling while presenting experimental results in English composed a challenging task. In addition, having the capability of combining different disciplines, such as statistics, language analysis and human communication, showed once more that Data Science is a multidisciplinary field.

Legacy

The descriptive approach is not as deeply addressed as the predictive one on studies related to NLP. Taking into account this context, there is a gap from the computational perspective, the understanding of how certain types of texts are characterised. As a consequence, it seems more arduous for machines to recognise than generate humorous texts [21].

In order to understand better how these two topics relate, it remains necessary an analysis regarding their characterisation. Furthermore, the social psychology field can benefit from these types of research because it could give orientation to intervention programmes to promote more respectful and healthy communication on social media. A research as the one presented in this work, can help to detect hurtfulness within social media, while being in figurative form (i.e. jokes). It is important to remark the fact that social ideas, and possible conflicts could be detected knowing a priory type of language used when are addressed in a humorous way. Data and code for developing this final project can be found in `https://github.com/lumer1/Final-Degree-Project.git`.

Data Science and NLP

Unstructured data is a type of information which is exponentially growing by the use of social media, surveys, crawlers, etc. These generate high volumes of data that must be processed (Figure 6.1). By improving the processing of unstructured information, it will be possible for companies and governments to get advantage and improve their business. As human language is something intrinsic in our everyday life, Data Science must supply the existent gap between computers and language by the aim of NLP.

Moreover, the detection of humour and offensive humour as well as their characterisation can help to improve processes of companies (e.g. opinion mining), while being useful to prevent hurtful messages disguised as jokes regarding any topic (women, immigrants, diseases) and help to detect potential dangers against social groups (e.g. minorities).

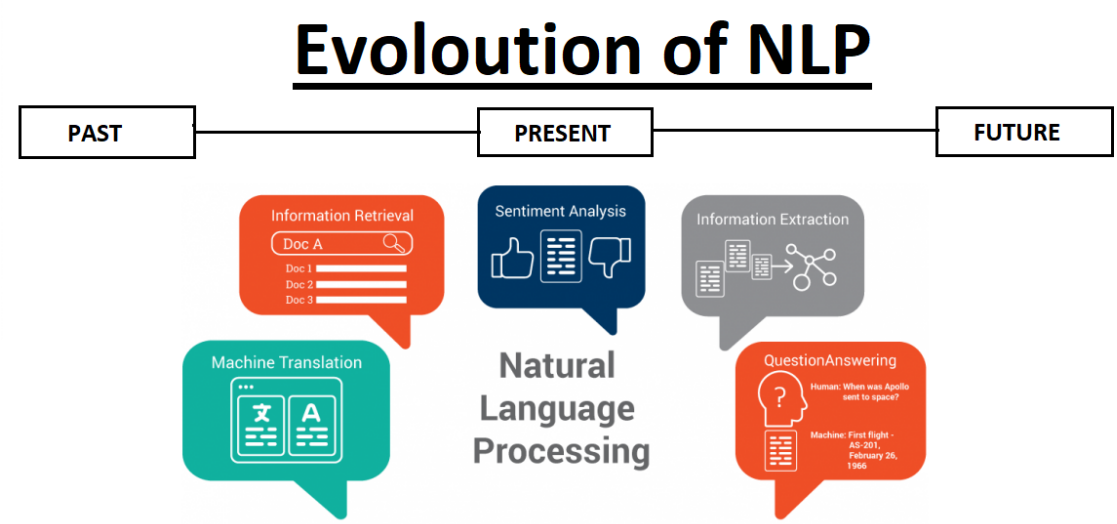


Figure 6.1: NLP applications [27]

6.3 Future Work

To address an improvement in this task, it would be interesting to apply a deeper tuning of the models. Hence, the challenge would consist in improving offensive humour recognition. A possible approach could be training the classification models with a set of linguistic features. Afterwards, it could be done an ablation test, with the most significant linguistic features that have the highest average of occurrence through offensive texts. Also, in line with the previous point, an implementation of other types of text representation such as Continuous Bag-of-Words, skip-grams and Global Vectors (*GloVe*) [28], might be relevant for this task.

A method to improve the detection can rely on data augmentation and subsequently, collecting hard cases of study, e.g. offensive jokes classified as non offensive. Afterwards, by carrying out an analysis of these texts, distinguishing if there is a recurring topic or relevant feature on them, in order to identify patterns of this type of jokes.
On the other hand, a research on the association that offensive language has with other types of hurtful language (hate-speech, stereotypes, aggressiveness, irony and sarcasm), could be applied to detect which are the most appealing terms for disguising hate within jokes.

Taking into account the novelty of the topic of this study, considering additional languages, will help to investigate if offensive humour differs with respect to the most common targets in each place. This should help to reach out easier ways to identify hate speech on social media.

Bibliography

[1] Joke example. https://twitter.com/nicoleosinga_rd/status/1268166158009671681?lang=bg. Accessed: 2022-06-27.

[2] What's the deal with accuracy, precision, recall and f1? https://towardsdatascience.com/whats-the-deal-with-accuracy-precision-recall-and-f1-f5d8b4db1021. Accessed: 2022-07-22.

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

[4] Elisa Bassignana, Valerio Basile, and Viviana Patti. Hurtlex: A multilingual lexicon of words to hurt. In *CLiC-it*, 2018.

[5] Arthur A Berger. What makes people laugh? cracking the cultural code. *Etc.*, 1975.

[6] Erik Cambria, Soujanya Poria, Rajiv Bajpai, and Bjoern Schuller. 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, December 2016. The COLING 2016 Organizing Committee.

[7] Francesca M M Citron, Marcus A Gray, Hugo D Critchley, Brendan S Weekes, and Evelyn C Ferstl. Emotional valence and arousal affect reading in an interactive way: neuroimaging evidence for an approach-withdrawal framework. *Neuropsychologia*, 2014.

[8] Jorge Carrillo de Albornoz, Laura Plaza, and Pablo Gervás. 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, May 2012. European Language Resources Association (ELRA).

[9] Andrea Esuli and Fabrizio Sebastiani. Sentiwordnet: A publicly available lexical resource for opinion mining. In *SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining*, 05 2006.

[10] Roushan Kumar Giri, Subhash Chandra Gupta, and Umesh Kumar Gupta. An approach to detect offence in memes using natural language processing (nlp) and deep learning. In *2021 International Conference on Computer Communication and Informatics (ICCCI)*, pages 1–5. IEEE, 2021.

[11] Janet Holmes and Meredith Marra. Humor and leadership style. *Humor - International Journal of Humor Research*, 19(2):119–138, 2006.

[12] Offensive humour graph. https://www.reddit.com/r/changemyview/comments/1roh7c/i_believe_that_a_joke_is_only_offensive_if_it_is/. Accessed: 2022-06-28.

[13] C. Hutto and Eric Gilbert. 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, May 2014.

[14] Internet growth statistics 1995 to 2022. https://www.internetworldstats.com/emarketing.htm. Accessed: 2022-07-01.

[15] Tonglin Jiang, Hao Li, and Yubo Hou. Cultural differences in humor perception, usage, and implications. *Frontiers in Psychology*, 10, 2019.

[16] S. G. Kwak and J. H. Kim. Central limit theorem: the cornerstone of modern statistics. *Korean J Anesthesiol*, 70(2):144–156, Apr 2017.

[17] Jake Lever, Martin Krzywinski, and Naomi Altman. Classification evaluation. *Nature Methods*, 13(8):603–604, Aug 2016.

[18] Esther Linares. La conceptualización lingüística del tabú en el discurso humorístico subversivo. *e-Scripta Romanica*, 7:79–96, 12 2019.

[19] Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60, 2014.

[20] J. Meaney, Steven Wilson, Luis Chiruzzo, Adam Lopez, and Walid Magdy. 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, 01 2021.

[21] Rada Mihalcea and Carlo Strapparava. 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, October 2005. Association for Computational Linguistics.

[22] Shazia Mirza and David Baddiel. Stand up to taboos: How comedy tackles the no-go subjects. *Index censorsh.*, 44(4):8–10, December 2015.

[23] Donna L. Mohr, William J. Wilson, and Rudolf J. Freund. Chapter 14 - nonparametric methods. In Donna L. Mohr, William J. Wilson, and Rudolf J. Freund, editors, *Statistical Methods (Fourth Edition)*, pages 651–683. Academic Press, fourth edition edition, 2022.

[24] Abhinav Moudgil. Short jokes. https://www.kaggle.com/datasets/abhinavmoudgil95/short-jokes, Feb 2017.

[25] Roberto Navigli and Simone Paolo Ponzetto. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250, 2012.

[26] Finn Årup Nielsen. A new ANEW: evaluation of a word list for sentiment analysis in microblogs. In Matthew Rowe, Milan Stankovic, Aba-Sah Dadzie, and Mariann 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, May 2011.

[27] Suneel Patel. NLP Guide 101: NLP Evolution — Past, Present and Future. https://suneelpatel18.medium.com/nlp-guide-101-nlp-evolution-past-present-and-future-fcc573629da3, 2020. Accessed: 2022-06-28.

[28] Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics.

[29] Soujanya Poria, Erik Cambria, Grégoire Winterstein, and Guang-Bin Huang. Sentic patterns: Dependency-based rules for concept-level sentiment analysis. *Knowledge-Based Systems*, 69:45–63, 2014.

[30] Antonio Reyes, Martin Potthast, Paolo Rosso, and Benno Stein. Evaluating humour features on web comments. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May 2010. European Language Resources Association (ELRA).

[31] Klaus R. Scherer. What are emotions? and how can they be measured? *Social Science Information*, 44(4):695–729, 2005.

[32] SenticNet. Sentic computing. https://sentic.net/computing/, 2022.

[33] U. Shoaib, N. Ahmad, P. Prinetto, and G. Tiotto. Integrating multiwordnet with italian sign language lexical resources. *Expert Systems with Applications*, 41(5):2300–2308, 2014.

[34] Jonas Sjöbergh and Kenji Araki. Recognizing humor without recognizing meaning. In Francesco Masulli, Sushmita Mitra, and Gabriella Pasi, editors, *Applications of Fuzzy Sets Theory*, pages 469–476, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg.

[35] Carlo Strapparava and Alessandro Valitutti. WordNet affect: an affective extension of WordNet. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal, May 2004. European Language Resources Association (ELRA).

[36] Yla Tausczik and James Pennebaker. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of Language and Social Psychology*, 29:24–54, 03 2010.

[37] Textblob: Simplified text processing. https://textblob.readthedocs.io/en/dev/, 2020.

[38] Tony Veale, Kurt Feyaerts, and Geert Brône. The cognitive mechanisms of adversarial humor. *Humor-international Journal of Humor Research - HUMOR*, 19:305–339, 08 2006.

[39] Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior Research Methods*, 45(4):1191–1207, Dec 2013.

[40] Constanze Weth. Distinguishing syntactic markers from morphological markers. a cross-linguistic comparison. *Frontiers in Psychology*, 11:2082, 08 2020.

[41] Michael Wiegand, Josef Ruppenhofer, Anna Schmidt, and Clayton Greenberg. 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, June 2018. Association for Computational Linguistics.

[42] Wenjie Yin and Arkaitz Zubiaga. Hidden behind the obvious: misleading keywords and implicitly abusive language on social media. In *Hidden behind the obvious: misleading keywords and implicitly abusive language on social media*, 05 2022.

Appendix A

Humour Detection. Tables with results.

Table A.1: Significant features for humour detection I

			Humour		Non Humour	
Tool or Lexicon	**Feature**	**p-value**	**Mean**	**Variance**	**Mean**	**Variance**
LIWC	I	7.46E-158	0.0587	0.0047	0.0198	0.0022
Part-of-Speech	1st Person Singular	5.95E-119	0.0318	0.002	0.0111	0.001
LIWC	Plural Pronouns	1.27E-55	0.1162	0.0063	0.0847	0.0063
LIWC	Pronouns	1.43E-49	0.1669	0.0071	0.1342	0.0097
LIWC	Hear	4.42E-47	0.013	0.0008	0.0053	0.0005
LIWC	Cognitive Mechanism	4.43E-41	0.1267	0.007	0.1619	0.0101
LIWC	She/He	6.53E-41	0.0185	0.0014	0.0098	0.0011
LIWC	Achieve	4.70E-40	0.0168	0.001	0.0305	0.0022
LIWC	Family	6.17E-38	0.0107	0.0007	0.0039	0.0003
LIWC	Positive Emotions	1.19E-36	0.0293	0.0019	0.0486	0.0043
LIWC	Article	2.99E-36	0.0805	0.0041	0.06	0.0033
LIWC	Past	2.98E-34	0.0364	0.0028	0.0201	0.0014
LIWC	Prepositions	3.95E-32	0.0966	0.0038	0.1175	0.0052
Hurtlex	Male Genitalia	5.63E-32	0.0287	0.0016	0.0179	0.0011
LIWC	Affective Processes	5.03E-31	0.0564	0.0037	0.0779	0.0061
Part-of-Speech	Pronouns	1.36E-30	0.0673	0.0022	0.0547	0.0023
EmoSenticNet	Joy	2.25E-30	0.3125	0.013	0.3481	0.0138
LIWC	Insight	8.06E-30	0.0147	0.0008	0.0251	0.0016
LIWC	We	8.07E-29	0.0042	0.0003	0.0127	0.0012
LIWC	Swear Words	1.39E-25	0.005	0.0004	0.0014	0.0001
Hurtlex	Negative Stereotypes and Ethnic Slurs	1.53E-22	0.0043	0.0003	0.0007	3.39e-05
EmoSenticNet	Surprise	4.81E-21	0.0482	0.0041	0.0297	0.0017
Part-of-Speech	1st Person Plural	2.65E-20	0.0022	0.0001	0.0065	0.0005
SentiSense	Like	5.47E-18	0.0261	0.0015	0.0362	0.0025
LIWC	Quantifiers	4.16E-16	0.0222	0.0013	0.0302	0.0018
SentiSense	Anticipation	2.08E-15	0.0124	0.0007	0.0184	0.0011
LIWC	Certainty	2.68E-14	0.0113	0.0006	0.0182	0.0013
Part-of-Speech	Punctuation Symbol	7.53E-14	0.1119	0.0041	0.1061	0.0068
LIWC	Inhibition	1.09E-13	0.005	0.0003	0.0085	0.0006
SentiSense	Love	1.23E-13	0.0051	0.0003	0.0096	0.0007
LIWC	Ingestion	1.08E-12	0.0135	0.0011	0.0088	0.0009
LIWC	Inclusive	2.95E-12	0.0299	0.0016	0.0386	0.0023
LIWC	Home	1.34E-11	0.0089	0.0006	0.0049	0.0003
LIWC	Friend	3.19E-11	0.0046	0.0003	0.0024	0.0002
Part-of-Speech	2nd Person Singular	6.12E-11	0.0008	0.0001	0.0028	0.0002
LIWC	Sad	1.22E-10	0.006	0.0004	0.0097	0.0008
LIWC	Tentative	8.91E-10	0.0173	0.001	0.0237	0.0016
Hurtlex	Female Genitalia	1.41E-09	0.0014	0.0001	0.0004	2.74e-05
LIWC	Anger	1.10E-08	0.0114	0.0007	0.0082	0.0007
LIWC	Perceptual Processes	1.81E-08	0.0364	0.0025	0.0308	0.0024

Table A.2: Significant features for humour detection II

			Humour		Non Humour	
Tool or Lexicon	**Feature**	**p-value**	**Mean**	**Variance**	**Mean**	**Variance**
LIWC	Sexual	7.16E-08	0.0114	0.0009	0.0079	0.0007
LIWC	Social	1.18E-07	0.1357	0.0092	0.1241	0.0094
LIWC	You	1.40E-07	0.0255	0.0018	0.0353	0.0031
LIWC	Humans	1.96E-07	0.0186	0.0011	0.0146	0.001
LIWC	Anxiety	9.42E-07	0.003	0.0002	0.0054	0.0004
LIWC	Function	1.23E-06	0.5208	0.0099	0.5026	0.0182
LIWC	Leisure	2.04E-06	0.0168	0.0012	0.0138	0.0012
LIWC	Verb	2.09E-06	0.1475	0.0046	0.1412	0.0072
Hurtlex	Animals	2.23E-06	0.0039	0.0003	0.0021	0.0001
EmoSenticNet	Anger	4.64E-06	0.0117	0.0007	0.0144	0.0008
LIWC	Relativity	5.07E-06	0.1187	0.0078	0.1314	0.0095
LIWC	Body	1.32E-05	0.0164	0.0012	0.0134	0.0011
LIWC	Health	1.58E-05	0.0124	0.0009	0.0162	0.0013
EmoSenticNet	Sad	2.12E-05	0.0438	0.0026	0.051	0.0034
LIWC	Adverb	5.46E-05	0.0417	0.0022	0.0386	0.0024
Part-of-Speech	Adjectives	7.50E-05	0.0901	0.0046	0.0973	0.0052
Hurtlex	Physical Disabilities and Diversity	8.20E-05	0.0025	0.0001	0.0016	0.0002
LIWC	Religion	0.00017	0.0071	0.0008	0.0043	0.0004
EmoSenticNet	Fear	0.0002	0.0116	0.0008	0.0144	0.0009
LIWC	Feel	2.51E-04	0.007	0.0004	0.009	0.0006
SentiSense	Disgust	2.84E-04	0.0065	0.0003	0.0089	0.0005
LIWC	Future	2.87E-04	0.032	0.0018	0.0274	0.0016
LIWC	Time	0.0005	0.0501	0.0034	0.057	0.0043
LIWC	Auxiliary verb	6.66E-04	0.093	0.0031	0.0903	0.0045
LIWC	Work	1.00E-03	0.0526	0.003	0.0582	0.0036
Part-of-Speech	Adverbs	1.08E-03	0.0506	0.0024	0.0495	0.0032
LIWC	I pronoun	1.16E-03	0.0045	0.0003	0.003	0.0002
SentiSense	Sadness	0.001	0.0154	0.0011	0.0182	0.0015
Hurtlex	Words Social and Economic Disadvantage	3.38E-03	0.0013	0.0001	0.0023	0.0002
LIWC	Biology	5.18E-03	0.0464	0.0034	0.0442	0.0042
LIWC	Death	6.62E-03	0.0057	0.0005	0.0037	0.0003
Part-of-Speech	3rd Person Singular	8.69E-03	0.0121	0.0007	0.0112	0.0008
LIWC	Nonfluencies	1.23E-02	0.002	0.0001	0.0012	0.0001
Hurtlex	Plants	1.35E-02	0.0026	0.0002	0.0014	0.0001
LIWC	Assent	1.71E-02	0.0059	0.0004	0.0054	0.0006
Hurtlex	Felonies crime and Immoral behaviour	1.88E-02	0.0033	0.0002	0.0037	0.0002
LIWC	Conjunction	2.03E-02	0.0462	0.0021	0.0494	0.0023
SentiSense	Joy	3.16E-02	0.0073	0.0004	0.0087	0.0008
LIWC	They	0.03	0.0094	0.0007	0.0072	0.0005

Appendix B

Offensive Humour Detection. Tables with results.

Table B.1: Significant features for offense detection within humour I using different tools.

			Non Offense		High Offense	
Tool or Lexicon Name	**Feature**	**p-value**	**Mean**	**Variance**	**Mean**	**Variance**
Lexicon Abusive words	Binary lexicon	6.16e-08	0.0003	1.74e-5.0	0.0024	0.0002
EmoSenticNet	Disgust	0.000368	0.0093	0.0006	0.0136	0.0009
	Surprise	2.139e-13	0.0409	0.0032	0.0639	0.0057
SentiSense	Anger	0.0014	0.0017	0.0001	0.0009	0.0001
	Anticipation	0.029	0.0133	0.0007	0.0119	0.0007
	Disgust	0.00030	0.03	0.002	0.0366	0.0022
	Fear	2.31e-11	0.0078	0.0005	0.015	0.001
	Joy	0.018	0.0058	0.0003	0.0094	0.0006
	Like	0.018	0.0276	0.0016	0.0244	0.0015
	Love	0.0274	0.0061	0.0003	0.0044	0.0003
	Sadness	0.0108	0.0033	0.0002	0.0053	0.0004
Part-Of-Speech	Noun	8.87e-05	0.2511	0.0088	0.2379	0.0092
	Adverbs	2.91e-05	0.0566	0.0033	0.048	0.0031
	Adjective	1.87e-07	0.0816	0.004	0.0968	0.0049
	1st Person Plural	3.25e-06	0.0033	0.0002	0.0012	0.0001
	1st Person Singular	1.72e-33	0.0383	0.0023	0.0183	0.0013
	2nd Person Singular	0.00493	0.0013	0.0001	0.0005	3.0e-5.0
	3rd Person Singular	0.00466	0.0123	0.0007	0.01	0.0006
Hurtlex	Negative Stereotypes & Ethnic-slurs	8.64e-40	0.0004	2.2e-5.0	0.0105	0.0008
	Moral & Behave Defects	2.56e-23	0.0023	0.0001	0.01	0.0006
	Felonies, Crime & Immoral Behave	0.0235	0.0028	0.0002	0.0039	0.0003

Table B.2: Significant features for offense detection within humour II with LIWC.

		Non Offense		High Offense	
Feature	**p-value**	**Mean**	**Variance**	**Mean**	**Variance**
Achieve	0.00234	0.0177	0.0011	0.0142	0.0009
Adverb	0.030	0.0426	0.0023	0.0387	0.0022
Affective Processes	0.00027	0.0583	0.004	0.0487	0.0031
Anger	0.00167	0.0087	0.0005	0.0127	0.0009
Anxiety	0.000165	0.0039	0.0002	0.0027	0.0002
Article	9.58e-10	0.0748	0.0038	0.0915	0.0047
Auxiliary Verb	2.67e-06	0.0902	0.003	0.1007	0.0031
Biology	1.089e-14	0.0363	0.0031	0.0534	0.0038
Body	2.72e-05	0.0136	0.0011	0.0178	0.0011
Cause	0.000634	0.0199	0.0015	0.0277	0.0022
Certainty	0.0006	0.0126	0.0007	0.0101	0.0006
Conjunction	0.0159	0.0432	0.0021	0.0469	0.0021
Discrepancy	0.0001	0.0145	0.0009	0.02	0.0012
Exclusive	1.88e-08	0.0213	0.0012	0.0143	0.0009
Feel	0.0082	0.0083	0.0005	0.0065	0.0004
Future	0.010	0.0071	0.0004	0.0051	0.0003
Health	0.00034	0.0088	0.0008	0.0132	0.001
Home	0.000346	0.0108	0.0009	0.0073	0.0005
Humans	2.11e-38	0.0103	0.0006	0.0283	0.0017
I	1.35e-45	0.0706	0.0051	0.0351	0.0036
Inhibition	0.00014	0.0056	0.0003	0.0036	0.0002
Insight	0.00022	0.0164	0.0009	0.0126	0.0007
Leisure	2.626e-07	0.0201	0.0015	0.0136	0.001
Motion	0.005	0.0206	0.0012	0.0174	0.0011
Number	0.032	0.0073	0.0005	0.0102	0.0007
Past	0.014	0.0356	0.0025	0.0335	0.0028
Perception	0.0012	0.0348	0.0025	0.0406	0.0027
Positive Emotions	2.38e-10	0.0322	0.0021	0.0223	0.0015
Personal Pronouns	4.605e-18	0.122	0.0064	0.0957	0.0061
Prepositions	3.867e-07	0.1037	0.0043	0.0893	0.0035
Present	3.125e-05	0.0948	0.0045	0.1049	0.0044
Pronoun	7.825e-13	0.1726	0.0076	0.1487	0.0067
Quantifiers	1.38e-07	0.0206	0.0011	0.0304	0.0019
Relativity	6.24e-06	0.1302	0.009	0.1114	0.0072
Religion	5.859e-15	0.0026	0.0002	0.0115	0.0012
See	1.226e-09	0.0111	0.0008	0.0197	0.0015
Sexual	2.388e-38	0.0038	0.0002	0.0198	0.0016
She/He	0.00083	0.0146	0.0011	0.0194	0.0014
Social	1.380e-12	0.1161	0.0088	0.1418	0.0091
Space	0.00084	0.0588	0.0033	0.0513	0.0031
Swear Words	6.95e-27	0.0009	5.0e-05	0.0082	0.0006
Tentative	0.0017	0.0178	0.001	0.015	0.0009
They	2.76e-07	0.0064	0.0004	0.0127	0.001
Time	0.0023	0.0542	0.004	0.0465	0.0032
We	2.861e-05	0.0052	0.0004	0.0026	0.0002
Work	0.00018	0.0163	0.0011	0.0129	0.001

Appendix C

Classification Tasks Metrics. Tables with results.

Table C.1: SVM with all features

Groups	**Precision**	**Recall**	**F1-score**	**Observations in Test set**
No offense	0.7	0.83	0.76	239
High offense	0.81	0.67	0.73	254
Macro average	0.75	0.75	0.75	493
Mean F1-Score with CV in train set	0.77			
Mean F1-Score with CV in test set	0.74			
Accuracy	0.74			

Table C.2: RF with all features

Groups	**Precision**	**Recall**	**F1-score**	**Observations in Test set**
No offense	0.71	0.81	0.76	239
High offense	0.79	0.69	0.74	254
Macro average	0.75	0.75	0.75	493
Mean F1-Score with CV in train set	0.746			
Mean F1-Score with CV in test set	0.732			
Accuracy	0.76			

Table C.3: LR with all features

Groups	**Precision**	**Recall**	**F1-score**	**Observations in Test set**
No offense	0.7	0.81	0.75	239
High offense	0.79	0.69	0.73	254
Macro average	0.75	0.74	0.74	493
Mean F1-Score with CV in train set	0.76			
Mean F1-Score with CV in test set	0.74			
Accuracy	0.73			

Appendix D

Reflection on the relationship between the final project and SDGs.

Sustainable Development Goals	High	Medium	Low	Not Applicable
SDG 1. End of poverty.				x
SDG 2. Zero hunger.				x
SDG 3. Health and wellbeing.	x			
SDG 4. Quality education.	x			
SDG 5. Gender equality.	x			
SDG 6. Clean water and sanitation.				x
SDG 7. Affordable and non-polluting energy.				x
SDG 8. Decent work and economic growth.				x
SDG 9. Industry, innovation and infrastructures.				x
SDG 10. Reduction of inequalities.	x			
SDG 11. Sustainable cities and communities.				x
SDG 12. Responsible production and consumption.				x
SDG 13. Climate action.				x
SDG 14. Underwater life.				x
SDG 15. Life in terrestrial ecosystems.				x
SDG 16. Peace, justice and solid institutions.	x			
SDG 17. Partnerships for achieving goals.				x

This final project is related to Sustainable Development Goals (SDGs) 03, 04, 05, 10, and 16. Concerning Goal 03 ("Health and well-being"), it is relevant to highlight that the majority of offensive jokes are directed toward minority groups. This kind of humour is systematically made up of mockery and references to aspects that characterise each of these groups. Consequently, the mental health and general well-being of people who belongs to these groups can be visibly affected by being a constant target of these types of jokes. On the other hand, for Goal 04 ("Quality education"), this work contributes to raising awareness of the offensive nature of certain types of humour, which undoubtedly has a remarkable educational value for both students and their environment. About Goal 05 ("Gender equality"), it should be pointed out that a large number of offensive jokes are directed at women, so an analysis of the characteristics of these jokes contributes to a more egalitarian society in terms of gender. One of the characteristics which help to discriminate between offensive and non-offensive humour is the presence of words referring to sexuality. Generally, these terms are used in jokes related to women. In the same way, the research contributes to Goal 10 ("Reduction of inequalities"), as one of the most striking results is that offensive humour applies especially to ethnic minorities. Hence, the presence of negative stereotypes and insults directed at ethnic minorities is significantly higher in offensive humour. The same occurs for the appliance of religious terms, which is notably higher in offensive humour. Overall, the work contributes to Goal 16 ("Peace, justice and strong institutions"), since all actions aimed at reducing inequalities between human beings, as well as offensive rhetoric, contribute to pacifying our societies and developing fairer and stronger institutions. In short, we believe that to advance in the development of well-being (SDG 03), reduce inequalities (SDG 05 and 10), and develop fairer institutions (SDG 04 and 16), it is necessary to consider how hatred towards certain groups is transmitted subtly through humour, and that is precisely the aim of this work.

Objetivos de Desarrollo Sostenibles	Alto	Medio	Bajo	No Procede
ODS 1. Fin de la pobreza.				x
ODS 2. Hambre cero.				x
ODS 3. Salud y bienestar.	x			
ODS 4. Educación de calidad.	x			
ODS 5. Igualdad de género.	x			
ODS 6. Agua limpia y saneamiento.				x
ODS 7. Energía asequible y no contaminante.				x
ODS 8. Trabajo decente y crecimiento económico.				x
ODS 9. Industria, innovación e infraestructuras.				x
ODS 10. Reducción de las desigualdades.	x			
ODS 11. Ciudades y comunidades sostenibles.				x
ODS 12. Producción y consumo responsables.				x
ODS 13. Acción por el clima.				x
ODS 14. Vida submarina.				x
ODS 15. Vida de ecosistemas terrestres.				x
ODS 16. Paz, justicia e instituciones sólidas.	x			
ODS 17. Alianzas para lograr objetivos.				x

El presente TFG se relaciona con los objetivos de desarrollo sostenible (ODS) 03, 04, 05, 10 y 16. En relación con el objetivo 03 ("Salud y bienestar") es importante destacar que gran parte de los chistes ofensivos se encuentran dirigidos hacia grupos minoritarios. Esta clase de humor se conforma de manera sistemática por burlas y referencias hacia aspectos que caracterizan cada uno de estos colectivos. En consecuencia, la salud mental y bienestar general de las personas pertenecientes a los mismos se puede encontrar visiblemente afectado al ser un blanco constante de este tipo de chistes. Por otro lado, en relación al objetivo 04 ("Educación de calidad"), el trabajo contribuye a tomar conciencia de el carácter ofensivo de cierto tipo de humor, lo que sin duda tiene un importante valor educativo tanto para los estudiantes como para su entorno. Respecto al objetivo 05 ("Igualdad de género") hay que señalar que una buena parte de los chistes ofensivos van dirigidos a mujeres por lo que un análisis de las características de estos chistes contribuye a una sociedad más igualitaria en términos de género. De hecho, una de las características que sirven para discriminar entre humor ofensivo y un humor no ofensivo es la presencia de términos que hacen referencia a la sexualidad. Generalmente estos términos se usan en los chistes relacionados con las mujeres. De la misma manera, la investigación contribuye al objetivo 10 ("Reducción de las desigualdades"), pues uno de los resultados más destacables es que el humor ofensivo se aplica especialmente a las minorías étnicas. De ahí que la presencia de estereotipos negativos e insultos dirigidos a minorías étnicas sea significativamente superior en el humor ofensivo. Lo mismo ocurre con la utilización de términos religiosos, que es significativamente superior en humor ofensivo. En su conjunto, el trabajo contribuye al objetivo 16 ("Paz, justicia e instituciones sólidas"), dado que todas aquellas acciones dirigidas a reducir las desigualdades entre los seres humanos, así como la retórica ofensiva contribuyen a pacificar nuestras sociedades y a desarrollar instituciones más justas y sólidas. En resumen, consideramos que para avanzar en el desarrollo del bienestar (ODS 03), reducir las desigualdades (ODS 05 y 10) y para desarrollar instituciones más justas (ODS 04 y 16) es necesario plantearse como el odio hacia ciertos colectivos se vehicula de manera sutil a través del humor, y ese es precisamente el objetivo de este trabajo.