Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 1278–1285
Marseille, 11–16 May 2020

Corpora and Baselines for Humour Recognition in Portuguese

Hugo Gonçalo Oliveira[1,2], **André Clemêncio**[1,2], **Ana Alves**[1,3]
[1]CISUC, University of Coimbra
[2]DEI, University of Coimbra
[3]ISEC, Polytechnic Institute of Coimbra
hroliv@dei.uc.pt, adfc@student.dei.uc.pt, ana@dei.uc.pt

Abstract

Having in mind the lack of work on the automatic recognition of verbal humour in Portuguese, a topic connected with fluency in a natural language, we describe the creation of three corpora, covering two styles of humour and four sources of non-humorous text, that may be used for related studies. We then report on some experiments where the created corpora were used for training and testing computational models that exploit content and linguistic features for humour recognition. Obtained results helped us taking some conclusions about this challenge and may be seen as baselines for those willing to tackle it in the future, using the same corpora.

Keywords: corpora creation, humour recognition, processing of Portuguese, text classification, feature extraction

1. Introduction

Computational Humour (Binsted et al., 2006) is a field of Artificial Intelligence that uses computers for detecting, analysing or producing humour. The automatic recognition of verbal humour is a branch of this field, with obvious connections with Natural Language Processing, because recognising humour expressed in a language, is a sign of fluency on that language (Tagnin, 2005). This means that an artificial agent that uses natural language for communication purposes should be able to recognise humour for better handling different situations. For instance, a news aggregator should have the ability of filtering out humorous news; or a chatbot should recognise humorous interactions and change its response, possibly ignoring it, or answering with generated humour as well.

Despite some work for other languages, especially English, humour recognition has not been an area of interest for Portuguese, until recently. While developing a computational model for humour recognition in Portuguese, we decided to tackle two styles of humour and gathered some texts for validation purposes (Clemêncio et al., 2019). The current paper presents three new corpora created out of those texts, but including additional styles of negative examples (why-questions and proverbs), and with balancing concerns, not only between the classes of humour and non-humour, but also between different sources. For instance, the corpus covering more sources has the positive examples balanced between two styles of humour (one-liners and humorous headlines) and the negative between four styles of non-humour (general knowledge questions, why-questions, proverbs, news headlines).

We further describe three new experiments where we tackle humour recognition as a supervised learning task, using the created corpora for validation, training and evaluation. Again, we went further than in previous work and present results for the best methods in the new corpora, after a more exhaustive search for the best parameters with 10-fold cross validation, and tests in unseen evaluation data, also exploring some additional features. This helped us take some conclusions on the challenge of humour recognition in Portuguese, and on possible features to explore and their relevance, also analysed in a final experiment. For instance, using an SVM and both content and humour-relevant linguistic features, both one-liners and headlines were recognised with $F1 = 0.88$ when they were the only positive instances. For the more challenging corpus that covers both styles of humour, F1 was ten points lower. Another conclusion was that exploiting exclusively content features lead to better performances than humour-relevant features alone. Yet, the best results are obtained when combining both. The reported results set baselines for future work, hopefully to be improved. In order to enable more studies on humour recognition in Portuguese, our corpora are now publicly available, for anyone willing to tackle this challenge.

After this introduction, we make a brief review of related work on humour recognition and corpora used for this purpose. We then present the created corpora, its sources and data distribution. Before concluding, we describe the performed experiments and report on the obtained results.

2. Related Work

The task of humour recognition has mainly been addressed for English. Most studies in this field tackle it as a binary classification problem between humour and no-humour. Yet, although there are many styles of humour, to simplify the task, most authors focused on a single style, generally short jokes or one-liners collected from web sources.

Mihalcea and Strapparava (2006) point that it is much easier to collect non-humorous data to be used as negative instances in humour recognition tasks. They could create a corpus with 16,000 humorous one-liners in English, collected from the Web, while, towards the development of a model of humour recognition, much negative data was available. Therefore, four sets of negative examples were gathered, namely: news titles from Reuters; proverbs on the Web; sentences from the British National Corpus (BNC); and sentences from the Open Mind Common Sense project. They stress that the negative examples should be similar in structure and composition to the one-liners, otherwise we risk that the models learn to distinguish examples based on non-relevant features for humour, such as text length or vocabulary used. Mihalcea and Strapparava (2006)'s corpus can be seen as a benchmark for humour recognition in English and has been used by other authors (Yang et al., 2015; Liu et al., 2018).

Mihalcea and Pulman (2007) augment the previous corpus with humorous news articles, which are longer. Sjöbergh and Araki (2007) created a smaller corpus with 6,100 one-liners, also collected from web sources. Their non-humorous examples came from the BNC. Smaller corpora for English include a set of 195 knock-knock jokes (Taylor and Mazlack, 2004) and 200 humorous headlines (Bucaria, 2004).

A corpus of a different nature was created for humour and irony recognition in Twitter (Barbieri and Saggion, 2014). It contains 40,000 tweets labelled with the categories of Irony, Education, Humour and Politics, according to the presence of corresponding hashtags (#irony, #education, #humor, #politics).

Besides the traditional bag-of-words approach for text classification, authors have focused on specific features that might be good indicators of humour. Those include the presence of idiomatic expressions and other typical joke words (Sjöbergh and Araki, 2007); human-centered vocabulary (e.g., "I", "you") (Mihalcea and Pulman, 2007); repetition of sounds in rhymes and alliteration (Mihalcea and Strapparava, 2006; Yang et al., 2015); antonyms and slang (Mihalcea and Strapparava, 2006); syntactic features (Liu et al., 2018), including number of phrases by type (e.g., NP, VP) or dependency relations; ambiguity (Sjöbergh and Araki, 2007; Barbieri and Saggion, 2014; Yang et al., 2015); sentiment / polarity of words (Mihalcea and Pulman, 2007; Barbieri and Saggion, 2014; Yang et al., 2015); and, of course, incongruity (Yang et al., 2015), in this case approximated by the inverse of the average similarity of the words used, in a model of distributional semantics.

There are linguistic studies on humour covering examples in Portuguese (Tagnin, 2005) but we do not know of any attempt at humour recognition in this language, except for our previous work (Clemêncio et al., 2019), where we used only two types of negative examples, were not so exhaustive in the search for the best parameters, and only presented validation results.

Yet, despite the lack of work on humour recognition, there is work on related topics in Portuguese, such as irony detection in Twitter. For this purpose, (Carvalho et al., 2009) created a corpus of tweets expressing opinions about political entities, and (de Freitas et al., 2014) focused on tweets mentioning the expression "fim do mundo", meaning end of the world, a trending topic in the end of 2012. On the topic of computational humour, there is also work on the automatic generation of potentially humorous riddles (Gonçalo Oliveira and Rodrigues, 2018).

3. Corpora Creation

When working on humour recognition for Portuguese, corpora for validation and testing is required. Yet, due to the lack of readily available corpora of such kind, we created our own corpus, which covers not only positive instances (i.e., humorous texts), but also negative (i.e., non-humorous texts), thus enabling text classification tasks. Given the underlying subjectivity involved, our sources had to be as consensual as possible. Also, positive and negative instances should not be too different, except in the actual features that humans rely on for discriminating between humour and non-humour. Otherwise, classifiers trained on the corpus may learn to differentiate classes based on features that are irrelevant for humour, e.g., length or structure of the text, or non-relevant vocabulary differences.

With this in mind, and because there are not many sources of Portuguese text, easily accessible and that we know, for sure, to be humorous, we first sought collections of texts from admittedly humorous sources. We first came across the "*Anedotário Português*"[1], a collection of jokes in Portuguese. From this collection, we extracted 342 short question-answering jokes, often called one-liners. Another 358 jokes of a similar kind were gathered from the Facebook page "*O Sagrado Caderno das Piadas Secas*"[2] (The Sacred Notebook of the Dry Jokes), where short jokes are regularly posted and books have been edited from (Pinto et al., 2017). This constitutes the first set of positive instances, for which we had to find negative, structurally similar but not humorous. As it is easier to find non-humorous text, we used two different sources, namely: 700 general-knowledge questions and answers in the Portuguese part of the parallel corpus MultiEight-04 (Magnini et al., 2004); and 1,446 "why" and "how" questions and their not so scientific answers, in many domains, from the recently closed website "*Os Porquês*"[3] (The Whys).

Yet, a system for humour recognition should not be restricted to a single style of humour. Therefore, to complement the collected text in the one-liners style, we targeted another kind of short humorous texts: humorous headlines. More precisely, we collected about 2,000 headlines posted in the Twitter account of *Inimigo Público* (IP), a humorous supplement of the Portuguese newspaper *Público*, between December 2018 and February 2019. For these, the negative instances were also collected from two different sources, namely: about 2,000 headlines from the Portuguese newspaper *Público*, posted in its Twitter account during February 2019; and 1,617 Portuguese proverbs available in the scope of the project Natura[4].

Table 2 illustrates the collected data with a text from each source and its given label, humour (H) or no-humour (N).

In the creation of our corpora, we were also concerned by data balance. Since training with imbalanced data may result in classifiers that favour the most common class, we decided to balance our corpus between humorous and non-humorous instances. We further balanced the negative examples between their sources. This was achieved by undersampling our data, which means that we forced the instances of each class to be the same number as the class with less instances. For the styles for which we had more than the necessary instances, the selection of those to include in the corpora was random.

Following this, we created three different corpora: one-liners, headlines and all, distributed according to Table 1.

[1] https://ltpf.files.wordpress.com/2011/01/omaiscompleto-anedotc3a3c2a1ri.pdf

[2] https://www.facebook.com/CadernoDasPiadas/

[3] http://osporques.com/ (Last time online on May 2019)

[4] https://natura.di.uminho.pt/~jj/pln/proverbio.dic

Due to the under-sampling, the headlines corpus ends up being the largest (4,000 instances), larger than the one covering both styles of humour (2,800). This happens because we could only collect 700 one-liners, also making the one-liners corpus the smallest (1,400 instances).

Corpus	Positive	Negative
One-liners (1,400 inst.)	Anedotário (342) Caderno (358)	MultiEight (350) Porquês (350)
Headlines (4,000 inst.)	IP (2,000)	Público (1,000) Proverbs (1,000)
All (2,800 inst.)	Anedotário (342) Caderno (358) IP (700)	MultiEight (350) Porquês (350) Público (350) Proverbs (350)

Table 1: Corpora sources and distribution

4. Features and Baselines for Humour Recognition in Portuguese

The balanced corpora created were used in some experiments, with the main focus on the binary classification of text into humorous or not, using traditional machine learning algorithms for supervised learning. This was first done when exploiting lexical features, then other linguistic features that are relevant for humour, and finally both. For this purpose, each corpus was randomly split into two sets, one for validation and training (80%) and another for testing (20%), also balanced. Experimentation was performed with the scikit-learn (Pedregosa et al., 2011) Python library. This section reports on the obtained results and ends with a brief analysis of relevant features on each corpora, identified with a χ^2 test.

Among other conclusions, obtained results show that it is harder to recognise humour, independently of the style it is conveyed in, than when targeting a single style. Given that we only used traditional approaches, there is obviously room for improvement, which is why we see these results as baselines for future work.

4.1. Exploiting Content features

In the first experiments, we exploited only content features and tested different classification methods (Naive Bayes, SVM and Random Forest (RF), with default parameters), weighting schemes (frequency count and TF.IDF), n-gram intervals ($n = 1$, $n = 1, 2$, $n = 1, 2, 3$, $n = 2$, $n = 2, 3$), and number of features to use. After several experiments on the three validation sets, the configuration that more consistently led to the highest performance was based on a SVM with a linear kernel, applied to a TF.IDF vector with 1,000 features, covering unigrams, bigrams and trigrams ($n = 1, 2, 3$). Results of a 10-fold cross validation and testing of a model trained with this configuration are in Table 3, for the three corpora.

Despite relying exclusively on content features, results are surprisingly high, considering that humour is often not trivial, and goes deeper than the lexical level. The best performance is for the one-liners corpus, with testing F1=0.87, suggesting that humour is easier to identify in the question-answering format. Despite ranking second in the validation, the lowest F1 (0.75) in the test is for the third corpus, which covers two styles of humour and negative instances from four distinct sources. This is especially affected by the lower recall, caused by a higher rate of humorous examples not classified as such (false negatives). It is also the corpus where the drop of performance between validation and test is more pronounced, caused by a higher difference between all the covered instances.

Despite the care taken in the selection of the sources, lexical differences that could not be avoided might have also played a role on these results. Still, they were in line with similar work for English (Mihalcea and Strapparava, 2006), where accuracies between 96% (against news titles) and 77% (against BNC sentences) were achieved with a SVM when identifying humorous one-liners. The main difference is that they used more data (32,000 instances balanced between positive and negative) and considered a single source of humour against four different sources of no-humour (news titles, BNC sentences, proverbs, commonsense statements), each in an independent test, none including texts in the question-answer form.

4.2. Exploiting Humour-relevant features

Supported by the literature on the topic, besides content features, we extracted several humour-relevant linguistic features to be considered by the learned models. Given that we are working on Portuguese text, once the features to extract were identified, we explored available linguistic resources for this language on which we could rely for their extraction. We believe that, besides humour recognition, most of these features might also be useful for other tasks in Portuguese text, like irony detection or emotion recognition. Some (e.g., incongruity, out-of-vocabulary words) are alternative applications of the exploited language resources.

For extracting these features, pre-processing was first performed with Python's NLTK, improved for Portuguese (Ferreira et al., 2019). It included tokenization, for dividing the text into tokens; part-of-speech (PoS) tagging, for identifying the PoS of each word; lemmatization, for identifying the dictionary form of each word-PoS pair; and named entity recognition, for identifying named entities (NEs) and assigning them a suitable category. Features that resort to lexicons where entries are lemmatized (e.g., antonymy, ambiguity) are extracted with the lemmatized version of the text, while the others use the original tokens.

Next, we enumerate the humour-related features considered, their motivation, and how they were extracted:

- The presence of negative polarity is often associated with the presence of humour (Mihalcea and Pulman, 2007), so we extract three **sentiment**-related features: number of words with positive polarity (Polarity #1); number of words with negative polarity (Polarity #2); whether there are more negative or positive words (Polarity #3). Polarities were obtained from SentiLex-PT (Silva et al., 2012), a polarity lexicon for Portuguese, where words like *beleza* (beauty) or *inteligência* (intelligence) have a positive polarity and words like *engano* (mistake), *pobre* (poor) or *morrer* (to die) have a negative polarity.

Example	Source	Label
Que jogam quatro elefantes dentro de um Mini? Squash! (What do four elephants play inside a Mini? Squash!)	Anedotário	H
Qual é a língua menos falada no mundo? Língua Gestual (What is the least spoken language in the World? Sign Language)	Caderno	H
Quem foi o primeiro presidente dos Estados Unidos? George Washington. (Who was the first president of the United States? George Washington.)	MultiEight-04	N
Porque o riso é contagioso? Rir é saudável e contagia de boa disposição quem está por perto. (Why is laughter contagious? Laughing is healthy and provokes good mood on those around.)	Os Porquês	N
Operação da GNR na estrada para fiscalizar condutores que comem carne na sexta-feira santa. (GNR operation on the road to inspect drivers who eat meat on Good Friday.)	Inimigo	H
Ministério das Finanças ainda não recebeu pedidos de pré-reforma no Estado. (Ministry of Finance has not yet received pre-retirement claims in the State.)	Público	N
mais depressa se encontra um mentiroso que um coxo. (a liar is found faster than a lame.)	Proverbs	N

Table 2: Examples of humorous (H) and non-humorous (N) texts and their sources.

	10-fold cross validation			**Test**		
	One-liners	Headlines	All	One-liners	Headlines	All
Precision	0.97±0.03	0.91±0.02	0.96±0.03	0.89	0.81	0.83
Recall	0.94±0.03	0.83±0.02	0.83±0.04	0.84	0.83	0.68
F1	0.96±0.02	0.87±0.02	0.89±0.03	0.87	0.82	0.75

Table 3: Best results exploiting exclusively content features.

- As humour often resorts to **slang**, we count the number of words listed in the *Dicionário Aberto de Calão e Expressões Idiomáticas*[5] (Open Slang Dictionary). Such words include *vaca* (cow/bitch), *merda* (shit), or *cagar* (to shit), among many others.
- As humour may resort to the repetition of sounds, **alliteration** is approximated by regularities in writing, addressed here by four features, namely the number of occurrences of the most frequent character uni/bi/tri/tetra-grams, extracted with the *ngrams()* function of NLTK.
- As humour is often associated to the presence of contradictory / **antonym** ideas, we count the number of pairs of lemmas that are antonyms in at least two out of ten Portuguese lexical knowledge bases (Gonçalo Oliveira, 2018). Examples of antonym pairs include *covardia-coragem* (cowardice-courage), *alegre-triste* (happy-sad), or *piorar-melhorar* (worsen-improve).
- Humour may explore different senses of the same word, here covered by two features related to **ambiguity**: average number of senses per lemma (Ambiguity #1), and highest number of senses for a lemma (Ambiguity #2), according to OpenWordNet-PT (Paiva et al., 2012), a Portuguese wordnet, where words are grouped in synsets, according to their possible senses.
- New words are often made up towards a humorous effect, so we have a feature for **out-of-vocabulary words**: number of words not covered by the vocabulary of a pre-trained word2vec CBOW model of word embeddings for Portuguese (Hartmann et al., 2017), with 300-sized vectors.
- **Incongruity**, in the core of one of the most popular theories of humour, is approximated by two features: average similarity of all pairs of words, and the lowest similarity score between a pair of words, both computed on the previous model of word embeddings. Given that incongruity is related to unexpectedness for being out of place, the lowest the value of this feature, the higher the incongruity of the text should be.
- **Named Entities** are widely used in news headlines and in general-knowledge questions and answers, so it would be interesting to understand its impact in our problem. This is covered by 11 features: number of NEs per category, considering the 10 categories in the HAREM collection (Freitas et al., 2010). The last feature is a sum of all NEs.
- Humour may also resort to mental **images**, while other texts (e.g., news) tend to be more **concrete**. This is covered by two features: average value of imageability, and average value of concreteness of all words, according to the Minho Word Pool norms (Soares et al., 2017).

New classifiers were also learned from the aforementioned 27 features, extracted from each text. Again, we experimented different methods but, in this case, the best performance was only achieved with a SVM for the headlines corpus. In the other two, a Random Forest performed better. Table 4 shows the validation and test results of both methods in the three corpora.

One first note, regarding the lower performance when compared to the previous experiment, is that, at least for these corpora, it seems to be more fruitful to exploit the words used, instead of the proposed linguistic features. This is also in line with similar work for English (Mihalcea and Strapparava, 2006), where results dropped when considering

[5] http://natura.di.uminho.pt/~jj/pln/calao/calao.dic.txt

	10-fold cross validation			Test		
	One-liners	Headlines	All	One-liners	Headlines	All
			SVM			
Precision	0.73±0.03	0.82±0.02	0.74±0.03	0.73	0.83	0.64
Recall	0.79±0.09	0.83±0.03	0.60±0.04	0.77	0.79	0.45
F1	0.76±0.04	0.83±0.02	0.66±0.03	0.75	0.81	0.53
			Random Forest			
Precision	0.82±0.03	0.81±0.02	0.77±0.03	0.83	0.82	0.75
Recall	0.77±0.05	0.77±0.03	0.67±0.04	0.76	0.72	0.56
F1	0.79±0.04	0.79±0.02	0.72±0.02	0.80	0.76	0.64

Table 4: Performance when exploiting exclusively humour-relevant features.

only three humour-relevant features (alliteration, antonymy, slang), less than the 27 features we covered, which is nevertheless much lower than the 1,000 content features used by the classifiers described the previous section. Yet, it is worth noticing that the drop of performance between validation and test is less pronounced here, suggesting that these features generalise better, especially for the corpora with a single style of humour. In opposition to the previous experiment, the best performance was in the headlines corpus, using a SVM, but only a single point higher than the one-liners corpus, with a Random Forest.

4.3. Exploiting both types of feature

In an attempt to get the best out of both types of feature, we combined them, namely the TF.IDF-weighted vector with the 1,000 content features, selected in the initial experiment, and the 27 humour-relevant features. In this experiment, we used the classification methods with the best performance in the previous, namely SVM and Random Forest, with results in Table 5 for the three corpora.

Results suggest that, when both types of feature are combined, performance is generally better, especially for the SVM with the headlines, where the F1 in test increases by 6 points. Also, although the Random Forest achieves the highest precision in the three corpora, the best testing F1 is always achieved with a SVM, due to its higher recall. Again, results are in line with previous work for English (Mihalcea and Strapparava, 2006), where minor improvements were also achieved when combining content features with antonymy, alliteration and slang.

4.4. Feature Relevance Analysis

A final experiment analysed the most relevant features for discriminating between classes in our corpora, independently of the used classification methods. This might help with better understanding the problem, and provide other useful insights on how humour is verbally conveyed in Portuguese. Table 6 shows the 15 most relevant features for each corpora, according to a χ^2 test.

Even though better results were obtained when using exclusively content features than humour-relevant features, according to the χ^2 test, most of the relevant features are of the latter kind. Table 7 has numbered examples that illustrate occurrences of relevant features.

In the top-10 most relevant, tokens only appear in the third corpus, namely the words 'quem' and 'porque', which are both interrogative pronouns, frequent in questions. Most usages of the former are in general-knowledge questions (see example #1) and proverbs (example #2), both negative examples, while the latter is used in 'why' questions (example #3), but also in one-liners (example #4).

Out-of-vocabulary words is the most relevant feature in the one-liners corpus and in the all corpus, but does not appear in the top for the headlines corpus. This is explained by the presence of made-up words in the one-liners (examples #5, #6). The number of named entities (#NEs) is highly relevant in the three corpora, which makes sense, because named entities are used in the negative examples of the first corpus (general-knowledge questions) and of the second (headlines). Some entity classes are also relevant. Person is the third most relevant feature in the headlines and the second in all, because several headlines target people (example #7). Location and organisation appear in the top for the first and second corpus because many general-knowledge questions and headlines mention this class of entities (examples #8, #9). The number of occurrences of the most frequent character (char unigrams feature) also seems to play a relevant role, especially in the second corpus, where it is the most relevant feature. This suggests that alliteration, a trend to repeat the same sound, is indeed relevant, but we would have to look deeper to have a stronger conclusion. Another relevant feature in all corpora is ambiguity, a common feature in humour. Additional relevant features in each corpus worth mentioning are polarity, imageability and concreteness. In opposition to what was expected, positive polarity (Polarity #1) is more relevant than negative, possibly because many negative examples from the Porquês corpus use positive words (example #10). As for imageability and concreteness in the second corpus, our interpretation is that real headlines are more concrete, while humorous ones are less and resort more often to mental images.

Out of the top-10, but in the top-15, we highlight the presence of two tri-grams commonly used in questions, in the first corpus (examples #4, #11, #12); of the token 'Marcelo', the first name of the Portuguese president, in the second corpus, due to its presence in many headlines, but mostly in humorous ones (47 out of 49), for being a very sociable person and swimming frequently on the sea (examples #13, #14); the slang feature, which we were expecting to be more relevant in the one-liners corpus than in the headlines, but was otherwise, possibly due to the utilization of 'soft' slang in the proverbs (examples #15, #16); and the bigram *os alentejanos*, the name of the people that live in Alentejo, an area in the south of Portugal, about which there

	10-fold cross Validation			**Test**		
	One-liners	Headlines	All	One-liners	Headlines	All
SVM						
Precision	0.96±0.03	0.89±0.02	0.92±0.02	0.88	0.87	0.83
Recall	0.94±0.04	0.89±0.02	0.86±0.03	0.88	0.88	0.73
F1	0.95±0.02	0.89±0.01	0.89±0.02	0.88	0.88	0.78
Random Forest						
Precision	0.94±0.03	0.87±0.01	0.88±0.03	0.90	0.88	0.91
Recall	0.85±0.05	0.75±0.03	0.72±0.04	0.80	0.63	0.66
F1	0.90±0.03	0.80±0.02	0.80±0.03	0.85	0.73	0.76

Table 5: Performance when exploiting both types of feature.

One-liners	**Headlines**	**All**
Out-of-vocabulary	Char unigrams	Out-of-vocabulary
#NEs	#NEs	NE ‘Person’
Char unigrams	NE ‘Person’	Ambiguity #2
Polarity #1	Ambiguity #2	Polarity #1
Ambiguity #2	Char bigrams	#NEs
NE ‘Location’	NE ‘Organization’	Char unigrams
NE ‘Time’	NE ‘Location’	Token ‘*quem*’ (who)
Char bigrams	Imageability	Token ‘*porque*’ (why)
NE ‘Organization’	NE ‘Value’	NE ‘Organization’
Polarity #3	Concreteness	NE ‘Work’
NE ‘Work’	Token ‘*quem*’ (who)	Imageability
Token ‘*um*’ (a/one)	Char trigrams	3-gram ‘*que é que*’ (what)
3-gram ‘*porque é que*’ (why)	Slang	3-gram ‘*porque é que*’ (why)
Token ‘*porque*’ (why)	Token ‘*não*’ (no)	Token ‘*sabem*’ (know)
NE ‘Value’	Token ‘*porque*’ (why)	Concreteness
3-gram ‘*que é que*’ (what)	Token ‘*Marcelo*’	2-gram ‘*os alentejanos*’

Table 6: Most relevant features.

are many jokes, mainly due to their stereotype of being too slow (examples #4, #12). Other linguistic features that we extracted (e.g., incongruity, antonymy) because we thought would be useful in this task, are, apparently, not so relevant, at least in these corpora.

5. Concluding remarks

We have presented newly created corpora for humour recognition in Portuguese, balanced between two styles of humour (one-liners and humorous headlines) and non-humorous text with a similar length and structure. We have also reported on some experiments where those corpora were used for training and testing models for humour recognition, which explored content and other linguistic features, relevant to this challenge.

All the texts collected, as well as the balanced corpora, are available from `https://github.com/andreclemencio/Recognizing-Humor-in-Portuguese/`. The corpus files only contain one text per line, ending with a tab followed by a ‘H’ or a ‘N’, respectively for positive or negative examples.

We sincerely hope that this work is only an initial step to this interesting field targeting the Portuguese language. Future work, may use the same corpora and look at the results reported here as baselines to beat. Despite some results higher than initially expected (e.g., when using only content features), there is definitely room for improvement, especially for the corpus with two styles of humour. Reported results were obtained with traditional machine learning methods and, to some extent, it would be interesting to test more recent classification methods, such as deep neural networks. However, the corpus might not be large enough for such an approach, which further motivates this kind of experiment. We may devise its augmentation but, as mentioned earlier, it might not be that straightforward to collect a large amount of consensual examples of verbal humour. We have explored automatic alternatives for this, such as retrieving tweets using certain hashtags (e.g., #humor, #piada), or focusing on Twitter accounts of Portuguese comedians, but there is always a significant amount of non-humorous tweets (e.g., comedians simply advertising their shows), so this might not be the best source.

Acknowledgements

This work was partially supported by FCT’s INCoDe 2030 initiative, in the scope of the demonstration project AIA, “Apoio Inteligente a empreendedores (chatbots)”.

6. Bibliographical References

Barbieri, F. and Saggion, H. (2014). Automatic detection of irony and humour in Twitter. In Proceedings of 5th International Conference on Computational Creativity (ICCC), pages 155–162.

Binsted, K., Nijholt, A., Stock, O., Strapparava, C., Ritchie, G., Manurung, R., Pain, H., Waller, A., and O’Mara, D. (2006). Computational Humor. *IEEE Intelligent Systems*, 21(2):59–69.

#	Text	Source
1	*Quem pintou a "Guernica"? Picasso.* (Who painted the Guernica? Picasso.)	MultiEight
2	*quem ri por último, ri melhor* (who laughs last laughs best)	Proverbs
3	*Porque temos medo? Ter medo é sobretudo recear a aceitação.* (Why are we afraid? To be afraid is above all to fear acceptance.)	Porquês
4	*Porque é que os alentejanos semeiam alhos nas bermas das estradas? Porque o alho faz bem à circulação.* (Why do alentejanos sow garlic on the roadside? Because garlic is good for circulation.)	Anedotário
5	*O que é um ponto preto no microscópio? Uma Blacktéria.* (What is a black spot on the microscope? A blackteria)	Anedotário
6	*Que nome se dá a um cão mágico? – Labracadabrador.* (What do you call a magic dog? – Labracadabrador.)	Caderno
7	*Lebron James supera Michael Jordan na lista dos melhores marcadores da NBA.* (Lebron James outperforms Michael Jordan on the NBA top scorers list.)	Público
8	*OCDE considera Portugal o quinto país com menor nível de discriminação contra mulheres.* (OECD considers that Portugal is the fifth country with the lowest level of discrimination against women.)	Público
9	*Onde é a sede da Lindt & Sprüngli? Kilchberg.* (Where is the headquarters of Lindt & Sprüngli? Kilchberg.)	MultiEight
10	*Quais os benefícios das uvas passas? As uvas passas desidratadas pela exposição ao sol, são uma boa fonte de fibra...* (What are the benefits of raisins? Raisins are a good source of fiber...)	Porquês
11	*O que é que vai e vem, sem sair do lugar? A porta.* (What comes and goes without leaving its place? The door.)	Caderno
12	*O que é que os alentejanos chamam aos caracóis? Chamam-lhes animais irrequietos.* (What do alentejanos call snails? They call them restless animals.)	Anedotário
13	*Marcelo vai visitar a nado todas as ilhas da Grécia* (Marcelo will swim to visit all the Greek islands)	IP
14	*Marcelo foi levado por arrasto por buscas da PJ em IPSS que estava a visitar* (Marcelo was dragged by police searches on the institution he was visiting)	IP
15	*na primeira quem quer cai; na segunda cai quem quer; na terceira quem é parvo.* (the first time anyone can be deceived; in the second, only who wants is deceived; in the third, only who is silly.)	Proverbs
16	*o melão e a mulher conhecem-se pelo rabo.* (both the melon and the woman are known by their bottom.)	Proverbs

Table 7: Illustrative examples of relevant features.

Bucaria, C. (2004). Lexical and syntactic ambiguity as a source of humor: The case of newspaper headlines. *Humor*, 17(3):279–310.

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

Clemêncio, A., Alves, A., and Gonçalo Oliveira, H. (2019). Recognizing humor in Portuguese: First steps. In Proceedings of 19th EPIA Conference on Artificial Intelligence, EPIA 2019, Vila Real, Portugal, September 3-6, 2019, Part II, volume 11805 of *LNCS/LNAI*, pages 744–756. Springer, September.

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

Ferreira, J., Gonçalo Oliveira, H., and Rodrigues, R. (2019). Improving NLTK for processing Portuguese. In Symposium on Languages, Applications and Technologies (SLATE 2019), page In press, June.

Freitas, C., Carvalho, P., Gonçalo Oliveira, H., Mota, C., and Santos, D. (2010). Second HAREM: advancing the state of the art of named entity recognition in Portuguese. In Proceedings of 7th International Conference on Language Resources and Evaluation, LREC 2010, La Valleta, Malta, May. ELRA.

Gonçalo Oliveira, H. and Rodrigues, R. (2018). Exploring lexical-semantic knowledge in the generation of novel riddles in Portuguese. In Proceedings of the 3rd Workshop on Computational Creativity in Natural Language Generation, CC-NLG 2018, pages 17–25, Tilburg, The Netherlands, November. ACL Press.

Gonçalo Oliveira, H. (2018). A survey on Portuguese lexical knowledge bases: Contents, Comparison and Combination. *Information*, 9(2).

Hartmann, N., Fonseca, E., Shulby, C., Treviso, M., Rodrigues, J., and Aluísio, S. (2017). Portuguese word embeddings: Evaluating on word analogies and natural language tasks. In Proceedings of 11th Brazilian Symposium in Information and Human Language Technology, STIL.

Liu, L., Zhang, D., and Song, W. (2018). Exploiting syntactic structures for humor recognition. In Proceedings of 27th International Conference on Computational Linguistics, pages 1875–1883, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

Magnini, B., Vallin, A., Ayache, C., Erbach, G., Peñas, A., de Rijke, M., Rocha, P., Simov, K. I., and Sutcliffe, R. F. E. (2004). Overview of the CLEF 2004 Multilingual Question Answering track. In Multilingual Information Access for Text, Speech and Images, 5th Workshop of the Cross-Language Evaluation Forum (CLEF), Revised Selected Papers, volume 3491 of *LNCS*, pages 371–391. Springer.

Mihalcea, R. and Pulman, S. (2007). Characterizing humour: An exploration of features in humorous texts. In International Conference on Intelligent Text Processing and Computational Linguistics, pages 337–347. Springer.

Mihalcea, R. and Strapparava, C. (2006). Learning to laugh (automatically): Computational models for humor recognition. *Computational Intelligence*, 22(2):126–142.

Paiva, V., Rademaker, A., and Melo, G. (2012). OpenWordNet-PT: An open Brazilian wordnet for reasoning. In Proceedings of 24th International Conference on Computational Linguistics, COLING (Demo Paper).

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Pinto, P., Ramalhinho, J., and Castro, G. (2017). O Caderno das Piadas Secas – 500 Tentativas de ter graça. Manuscrito Editora.

Silva, M. J., Carvalho, P., and Sarmento, L. (2012). Building a sentiment lexicon for social judgement mining. In Proceedings of 10th International Conference on Computational Processing of the Portuguese Language (PROPOR 2012), volume 7243 of *LNCS*, pages 218–228, Coimbra, Portugal. Springer.

Sjöbergh, J. and Araki, K. (2007). Recognizing humor without recognizing meaning. In International Workshop on Fuzzy Logic and Applications, pages 469–476. Springer.

Soares, A. P., Costa, A. S., Machado, J., Comesaña, M., and Oliveira, H. M. (2017). The Minho Word Pool: Norms for imageability, concreteness, and subjective frequency for 3,800 Portuguese words. *Behavior Research Methods*, 49(3):1065—1081.

Tagnin, S. E. (2005). O humor como quebra da convencionalidade. *Revista brasileira de linguística aplicada*, 5(1):247–257.

Taylor, J. M. and Mazlack, L. J. (2004). Computationally recognizing wordplay in jokes. In Proceedings of the Annual Meeting of the Cognitive Science Society, pages 2166—2171, Stresa, Italy.

Yang, D., Lavie, A., Dyer, C., and Hovy, E. (2015). Humor recognition and humor anchor extraction. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 2367–2376. ACL Press.