From Humour to Hatred: A Computational Analysis of Off-Colour Humour

Vikram Ahuja(✉), Radhika Mamidi, and Navjyoti Singh

International Institute of Information Technology, Hyderabad, India
vikram.ahuja@research.iiit.ac.in, radhika.mamidi@iiit.ac.in

Abstract. Off-colour humour is a category of humour which is considered by many to be in poor taste or overly vulgar. Most commonly, off-colour humour contains remarks on particular ethnic group or gender, violence, domestic abuse, acts concerned with sex, excessive swearing or profanity. Blue humour, dark humour and insult humour are types of off-colour humour. Blue and dark humour unlike insult humour are not outrightly insulting in nature but are often misclassified because of the presence of insults and harmful speech. As the primary contributions of this paper we provide an original data-set consisting of nearly 15,000 instances and a novel approach towards resolving the problem of separating dark and blue humour from offensive humour which is essential so that free-speech on the internet is not curtailed. Our experiments show that deep learning methods outperforms other n-grams based approaches like SVM's, Naive Bayes and Logistic Regression by a large margin.

Keywords: Off-colour humour · Deep learning · Insult detection

1 Introduction

In the last decade, there has been an exponential increase in the volume of social media interactions (twitter, reddit, facebook etc.). It took over three years, until the end of May 2009, to reach the billionth tweet [1]. Today, it takes less than two days for one billion tweets to be sent. Social media has increasingly become the staple medium of communication via the internet. However, due to the non-personal nature of online communication, it presents a unique set of challenges. Social media has become a breeding ground for hate speech and insults as there is a lack of accountability that can be abused.

We need to discern between content which is an honest attempt at humour as opposed to content which is purely derogatory and insulting. Humour is an essential part of communication and allows us to convey our emotions and feelings. Humour is the tendency of particular cognitive experiences to provoke laughter and provide amusement.

Humour that is sometimes considered to be purely offensive, insulting or a form of hate speech is described as off-colour humour. Off-colour humour (also

M. Zhang et al. (Eds.): NLPCC 2018, LNAI 11109, pp. 144–153, 2018.
https://doi.org/10.1007/978-3-319-99501-4_12

known as vulgar humour, crude humour, or shock humour) is humour that deals with topics that may be considered to be in poor taste or overly vulgar. It primarily consists of three sub-categories, dark humour, blue humour and insult humour.

Dark Humour has been more frequently discussed in literary, social research as well as psychology but not much attention has been given to it in linguistics. It is a comic style that makes light of subject matter that is generally considered taboo, particularly subjects that are normally considered serious or painful to discuss such as death as defined by wikipedia. Dark humour aims at making fun of situations usually regarded as tragic, such as death, sickness, disability, and extreme violence, or of the people involved or subject to them [4]. It is inspired by or related to these tragic events and do not in any way make fun of them [5]. In dark humour a gruesome or a tragic topic is mixed with an innocuous topic which creates shock and inappropriateness. This invoked inappropriateness or shock generally amusing to the listeners [6]. Dynel [7] in their paper show that dark humour inspired by tragic events such as a terrorist attacks just addresses topics tangential to them and do not in any way make fun of them directly. Blue humour is a style of humour that is indecent or profane and is largely about sex. It contains profanity or sexual imagery that may shock. It is also referred to as Ribaldry. Insult humour is that kind of humour which consists of offensive insults directed to a person or a group. Roasting is a form of insult comedy in which specific individual, a guest of honor, is subjected to jokes at their expense, intended to amuse the event's wider audience as defined by Wikipedia. All the three categories mentioned above seem to be interrelated to each other but have very fine differences. Dark humour is different from straightforward obscenity (blue humour) in the way that it is more subtle. Both dark and blue humour are different from insult humour in the sense that there is no intent of offending someone in the former two whereas in insult humour the main aim is to jokingly offend or insult the other person or a group of people [8].

People often get offended on such misunderstood instances of humour more than would otherwise be the case. The significance of our contribution can be fully conceived only when we realise that such occurrences can lead to gratuitous censorship and therefore curtailment of free speech. It is in this context that we are trying to formulate our problem of separating dark humour and blue humour from insult humour. In short our contributions can be summarised as follows:

- We present a dataset of nearly 15,000 jokes out of which 4,000 are of positive types.
- A novel approach towards resolving the problem of separating dark and blue humour from offensive humour.

The remainder of the paper is structured as follows. Section 2 gives a detail about related work along with it's criticisms. Section 3 presents the proposed framework. Section 4 presents the dataset used, Sect. 5 presents the experiments used. Section 6 gives the results and analysis of the experiment conducted in this study and Sect. 7 concludes the paper (Table 1).

Table 1. Few examples of jokes used in our dataset

Dark joke	1. My girlfriend is a porn star. She will kill me if she finds out 2. When someone says, "Rape jokes are not funny," I don't care. It's not like I asked for their consent anyway
Blue joke	1. Sex is not the answer. Sex is the question. "Yes" is the answer 2. How can you tell if a man is sexually excited? He's breathing
Insult joke	1. Your mama so fat when she stepped on the weighing scale it said: "I need your weight, not your phone number." 2. You were beautiful in my dreams, but a fucking nightmare in reality
Normal/safe joke	1. What's red and bad for your teeth? A brick 2. What did the German air force eat for breakfast during WW2? Luftwaffle

2 Related Work

Humour has always been an important topic for researchers. There has been a lot of study in humour in the field of linguistics, literature, neuroscience, psychology and sociology. Research in humour has revealed many different theories of humour and many different kinds of humour including their functions and effects personally, in relationships, and in society. For the scope of this paper we are restricting ourselves to off-colour humour that has been explained in the above sections.

There has been some studies on offensive humour which is usually used a form of resistance in tragic situations. Weaver [10] in their paper talks how racism could be undermined by using racial stereotypes by blacks and minority ethnic comedians. Lockyer [11] in their study analyse how disabled comedians have also ridiculed stereotypes of the disabled by reversing the offensive comments of the non-disabled.

The study by Billig [12] examines the relationship between humour and hatred, which it claims is the topic that is often ignored by researchers of prejudice. It analyses websites that present racist humour and display sympathies with the Ku Klux Klan. The analysis emphasizes the importance of examining the *metadiscourse*, which presents and justifies the humour and also suggests that the extreme language of racist hatred is indicated to be a matter for enjoyment. In the book "Jokes and their Relation to the Unconscious", Freud refers to off-colour humour as the "economy of pity" and claims that it is "one of the most frequent sources of humourous pleasure" and these jokes (off-colour) provides a socially accepted means of breaking taboos, particularly in relation to sex and aggression. In [7] the authors analyse the importance of off colour humour (dark humour) in post terrorist attack discourse. The paper claims that dark humour is a coping mechanism under oppressive regimes and in crisis situations. Davies [14] in his book argues that those who engage in racist and sexist jokes

do not necessarily believe the stereotypes that the jokes express. Maxwell [15] in his paper brings forth the importance of dark humour as a cognitive and/or behavioral coping strategy which is considered to be a reaction to a traumatic event and proposes a model including progressive steps of humour, ranging from respectful to sarcastic.

Similarly there has been a lot of studies in the field of insult detection. Mahmud et al. [16] in their paper create a set of rules to extract the semantic information of a given sentence from the general semantic structure of that sentence to separate information from abusive language but is very limited in the sense that this system can annotate and distinguish any abusive or insulting sentence only bearing related words or phrases that must exist in the lexicon entry. So, it only looks at insulting words and not sentences that are used in an insulting manner.

Xiang et al. [17] in their work dealt with offensive tweets with the help of Topical Feature Discovery over a Large Scale Twitter Corpus by using Latent Dirichlet Allocation model. The work by Ravazi et al. [18] describe an automatic flame detection method which extracts features at different conceptual levels and applies multilevel classification for flame detection but is very limited due to the dataset used by them and does not consider the syntactical structure of the messages explicitly.

The above works tell us there has been quite a lot of studies in both these fields but no such computational study in the intersection of these two topics. This study is the first such attempt which tries to create a separating boundary between different types of off-colour humor and insults. We discuss our framework used to separate different types of off-colour humour in the next section.

3 Proposed Framework

The domain of jokes we are dealing with, viz. dark humour, blue humour and insult humour all are generally classified under the umbrella category of NSFW or Off-Colour Humour. It is due to their apparent similarities that on one glance they can be dismissed as being of one and the same type. As we go to finer levels of granularity it becomes evident that two separate buckets can be defined even inside off-colour humour, one pertaining to insults resulting in insult humour and the other consisting of dark humour and blue humour as shown in the Fig. 1 below. Insults being the common denominator does not mean all insults and non-jokes can be classified as insult humour, thus defining a demarkation separating the two is also required.

As mentioned above we are able to define clear boundaries between within off-colour humour between insulting and non-insulting humour. But in order to further differentiate between dark and blue humour we identify more features that can give us a clear distinction between the two. One of the primary indicators in helping us draw this line is the ability to detect and extract sexual terms leading us to blue humour and dark themes such as violence (murder, abuse, domestic violence, rape, torture, war, genocide, terrorism, corruption), discrimination (chauvinism, racism, sexism, homophobia, transphobia), disease (anxiety,

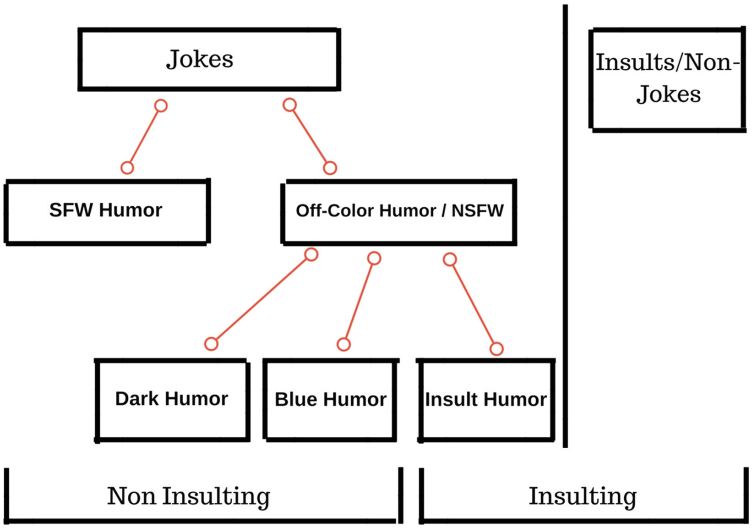

Fig. 1. Differentiating between Insulting and Non-Insulting Humor

depression, suicide, nightmares, drug abuse, mutilation, disability, terminal illness, insanity), sexuality (sodomy, homosexuality, incest, infidelity, fornication), religion and barbarism [2] leading us to dark humour. In order to define strong outlines we also ensure that even if an insulting joke or an insult contains sexual content or dark themes the primary focus of such content is on the insult part and not its sexual content or dark theme (which more than anything aides in providing a backdrop).

Jokes	Insult	Result
0	0	Non-Insulting Statement
0	1	Insulting Statement
1	0	Non Insulting Jokes
1	1	Insulting Jokes

Fig. 2. In this figure Dark and Blue Humor belongs to the category of Non-Insulting Jokes which are differentiated by dark and blue humor features mentioned in the above paragraph

The focus of this paper is to classify between these three categories of off-colour and hence, integrating the task of classification of text between humourous

and non-humourous has been deemed out of scope and is limited by the creation of a dataset which consists of only the humourous content.

4 Dataset

To test our hypothesis that automatic classification models can classify between dark humour, blue humour, Insult humour and (normal humour), we needed a dataset consisting of all types of above mentioned examples. Since there is no such corpus available for the task because of limited study in this field, we collected and labeled our own data. The only data source that was available is the twitter dataset [17] which has been used to detect offensive tweets but due to the fact it was very limited in terms of themes and could not have been used for our study.

The dataset that we created consisted of multiple one-liners. We used one-liners because they are very small generally and must produce humourous effect, unlike longer jokes which usually have a relatively complex narrative structure. These characteristics make this type of humour particularly suitable for use in an automatic learning setting. The dataset is defined as follows:

- **Insult jokes:** We collected multiple one-liners jokes from the subreddits /r/insults and /r/roastme. Apart from those we also mined various jokes websites and collected jokes with tags "Insult". After removing duplicates and verifying manually we were left with nearly 4000 jokes belonging to the category of insult jokes.
- **Dark Jokes:** We collected multiple jokes from the subreddit /r/darkjokes and /r/sickipedia. These subreddits contains one-liner jokes which are highly moderated. Apart from that we mined various jokes websites and collected jokes with the tags dark, darkjokes and darkhumour. After removing the duplicates and manual verification we were left with a final dataset of approximately 3500 jokes under the category of dark jokes.
- **Blue Jokes:** Blue jokes are the types of jokes which are most famous on the internet. Since these types of jokes are mainly associated with heavy nudity, sexual content and slangs we collected one liner jokes from subreddit /r/dirtyjokes and apart from that we took jokes from various jokes websites with tags NSFW, dirty, adult and sexual. After duplicates removal and manual verification we were left with approximately 2500 jokes under the category of blue jokes.
- **Normal Jokes/Safe jokes:** We collected jokes from subreddits r/cleanjokes and /r/ oneliners. These subreddits contain clean, non offensive jokes and non disrespectful jokes. Jokes in this category does not belong to any of the above category. This types of jokes are referred to as SFW (safe for work) jokes for all future references. After collecting these jokes we searched for insult words (talk about a lexical dictionary from the paper opened in one of the tabs). After duplicates removal we were left were approximately 5000 jokes under this category.

The dataset collected is important because of the fact that it is multidimensional in the sense that it contains insulting and non insulting jokes as well jokes with taboo topics (mentioned in the above section) as well as jokes without these topics. This leaves us with a combined total of nearly 4000 jokes under the category of insult jokes, 3500 under the category of dark jokes (non-insult), 2500 jokes in the category of blue jokes (non-insult) and 5000 jokes from the category of safe jokes (non insult). Thus we have a 4000 positive examples and 11,000 negative examples. All the dataset that has been mined has been taken from websites which have strict moderation policies, thus leaving very little room for error in our dataset.

5 Experiment

Treating the problem of separating different types of jokes as a classification problem, there are wide variety of methods that can be used which can greatly affect the results. Some of the features explicitly used were:

- Dark jokes are usually limited to or their main topic is violence (murder, abuse, domestic violence, rape, torture, war, genocide, terrorism, corruption), discrimination (chauvinism, racism, sexism, homophobia, transphobia), disease (anxiety, depression, suicide, nightmares, drug abuse, mutilation, disability, terminal illness, insanity), sexuality (sodomy, homosexuality, incest, infidelity, fornication), religion and barbarism. In order to detect these relations a common sense engine called "Concept Net" [19]. It is a multilingual knowledge base, representing words and phrases that people use and the common-sense relationships between them. Concept Net was used too see use of words related to the above mentioned topics.
- Sentiment score of every joke was calculated because of the hypothesis that off-colour humour tends to have a negative sentiment compared to jokes without any vulgar or any such topics mentioned above.
- It is our hypothesis that most of the insult jokes have mainly first (self-deprecating humour) and second (directed towards someone) person feature words like "I", "you", "your". This is done in order to detect insulting jokes with contains phrases like "your mother", "your father" which are usually meant to be insults.

After preprocessing of the data we extracted n-grams from the dataset, precisely speaking unigrams, bigrams and trigrams and a feature dictionary was created using of these collected ngrams. We compare results from five different classification algorithms mentioned below with different settings along with features mentioned above.

- We used LDA, which provides a transitive relationship between words such that each document has a set of words and each word belong to multiple topics (here categories of jokes) which transitively indicates that each document is a collection of topics.

– Used n-grams trained a logistic regression and naive bayes with it.
– Along with this we used brown clustering which helped us to put similar kinds of words in same cluster and trained a SVM with it.
– We also experimented with CNN's like Kim et al. [20] work which uses CNN for sentence classification. A model as shown in [20] was created with pre-trained vectors from *word2vec*, which has been trained on google news corpus and the vectors have dimensionality of 300.

Various experiments were performed on our dataset. For evaluation metrics, the dataset was randomly divided into 90% training and 10% testing. All the experiments were performed 10 fold and the final result was then taken to be average of those results.

6 Analysis

We can see the results of our classifiers in the Table 2 below. In the case of Logistic Regression, the introduction of features suggested proved to be an important factor as it increased the accuracy by nearly 10%. LDA had a slight better result than Logistic Regression without any features but is outperformed by Logistic Regression when the features such as sentiment scores, first, second person features and dark words are added. Naive bayes outperforms both LDA and Logistic Regression when features such as sentiment scores, first and second person features and dark word are not added, but we see equal results when those are added. SVM outperforms LDA, Naive bayes and Logistic Regression by a big margin and thus proving to be a better algorithm to classify. We see that the feature set used proved to be a very valuable addition to our experiment and an increase in accuracy in every case when those features are introduced. Finally, CNN's are introduced along with word2vec outperforms every other classifier used in this study. Thus we achieve the best accuracy of 81% using CNN's with *word2vec*.

Table 2. Table showing the accuracies of various classifiers used

Results	
Features	Accuracy
Logistic Regression	59%
LR + Ngrams	62%
LR + Ngrams + features mentioned	69%
LDA	61%
Naive Bayes + Ngrams + features mentioned	69%
SVM	68%
SVM + Ngrams + features	74%
CNN + word2vec	81%

7 Future Work

Given the constraints of the scope of the paper as well as the research conducted we have not attempted to integrate the differentiate humourous and non-humourous text in our study. This could also be incorporated in the pipeline to match with other studies. Also, in this paper we have restricted sexual topics in dark humour (not to be confused with sexuality in blue humour) but in reality, there are some dark jokes which have some common features with blue jokes or talks about nudity and profanity. This can be taken up for future work and this whole system could be implemented on various social media platforms to a more holistic classification to insults in social media and practice free speech.

References

1. numbers. Twitter Official Blog, 14 March 2011
2. Black Comedy. https://en.wikipedia.org/wiki/Black_comedy
3. Ribaldry. https://en.wikipedia.org/wiki/Ribaldry
4. Bucaria, C.: Dubbing dark humour: a case study in audiovisual translation. Lodz Pap. Pragmat. **4**(2), 215–240 (2008)
5. Lewis, P.: Three Jews and a blindfold: the politics of gallows humour. In: Avner, Z., Anat, Z. (eds.) Semites and Stereotypes: Characteristics of Jewish Humour, pp. 47–58. Greenwood Press, Westport (1993)
6. Aillaud, M., Piolat, A.: Influence of gender on judgment of dark and non-dark humour. Individ. Differ. Res. **10**(4), 211–222 (2012)
7. Dynel, M., Poppi, F.I.M.: In tragoedia risus, analysis of dark humour in post-terrorist attack discourse. Discourse Commun. (2018)
8. Kuipers, G.: 'Where was King Kong when we needed him?' Public discourse, digital disaster jokes, and the functions of laughter after 9/11 (2011)
9. Gournelos, T., Greene, V., (eds.): A Decade of Dark Humour: How Comedy, Irony and Satire Shaped Post-9/11 America, pp. 20–46. University Press of Mississippi, Jackson
10. Weaver, S.: Developing a rhetorical analysis of racist humour: examining anti-black jokes on the Internet. Soc. Semiot. **20**(5), 537–555 (2010). https://doi.org/10.1080/10350330.2010.513188
11. Lockyer, S.: From comedy targets to comedy-makers: disability and comedy in live performance. Disabil. Soc. **30**(9), 1397–1412 (2015). https://doi.org/10.1080/09687599.2015.1106402
12. Billig, M.: Humour and hatred: the racist jokes of the Ku Klux Klan. Discourse Soc **12**(3), 267–289 (2001)
13. Freud, S.: Jokes and their relation to the unconscious. WW Norton & Company (1960)
14. Davies, C.: Ethnic humour around the world: a comparative analysis. Indiana University Press (1990)
15. Maxwell, W.: The use of gallows humour and dark humour during crisis situations. Int. J. Emerg. Ment. Health **5**(2), 93–98 (2003)
16. Mahmud, A., Ahmed, K.Z., Khan, M.: Detecting flames and insults in text (2008)
17. Xiang, G., Hong, J., Rose, C.P.: Detecting offensive tweets via topical feature discovery over a large scale twitter corpus. In: Proceedings of the 21st ACM Conference on Information and Knowledge Management, Sheraton, Maui Hawaii, 29 October–2 November 2012 (2012)

18. Razavi, A.H., Inkpen, D., Uritsky, S., Matwin, S.: Offensive language detection using multi-level classification. In: Farzindar, A., Kešelj, V. (eds.) AI 2010. LNCS (LNAI), vol. 6085, pp. 16–27. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-13059-5_5
19. Liu, H., Singh, P.: Concept net - a practical commonsense reasoning tool-kit. BT Technol. J. **22**(4), 211–226 (2004)
20. Kim, Y.: Convolutional neural networks for sentence classification. In: EMNLP, pp. 1746–1751 (2014)